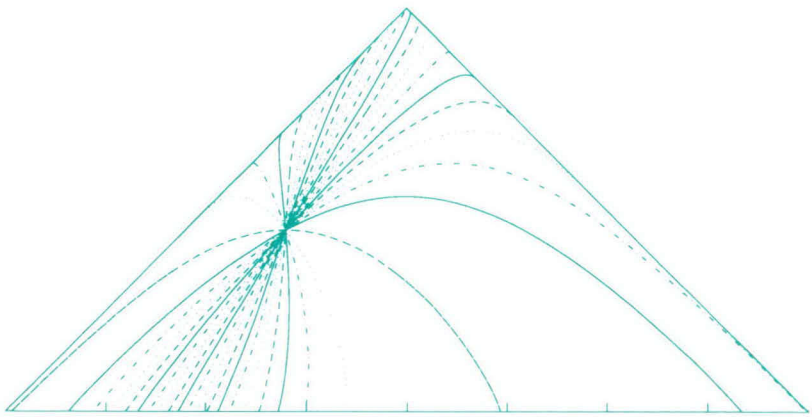
Differential geometric models for time-varying coefficients of autoregressive processes

Rijkeboer, Adriaan Lucas

Publication date:
1994

Document Version
Publisher's PDF, also known as Version of record

Differential Geometric Models for Time-varying Coefficients of Autoregressive Processes



Arie L. Rijkeboer

Differential Geometric Models for Time-varying Coefficients of Autoregressive Processes

Proefschrift ter verkrijging van de graad van doctor
aan de Katholieke Universiteit Brabant, op gezag
van de rector magnificus, prof. dr. L.F.W. de Klerk,
in het openbaar te verdedigen ten overstaan van
een door het college van dekanen aangewezen commissie
in de aula van de Universiteit op

vrijdag 29 april 1994 te 16.15 uur

door

Adriaan Lucas Rijkeboer

geboren te Arnhem.

Promotoren: Prof. Dr. J.M. Schumacher
Prof. Dr. M.H.C. Paardekooper

Co-promotor: Dr. J.C. Engwerda

Preface

For a long time I was a secondary school teacher of mathematics, but some years ago I wanted to try another profession. I am grateful to Tilburg University for enabling me, in my late thirties, to develop my abilities in the field of scientific research. I was lucky to get the guidance of three researchers, who provided me with many valuable suggestions, ideas and references. The intensive weekly discussions the four of us held were pleasant and very stimulating for my research. I thank my three advisors because they all did their very best to make my project a succes. Prof. Hans Schumacher put the fundamental question:'why should the coefficient space be regarded as linear in the model?' and this question was the starting point of the research leading to this thesis. He also gave important references and followed my work closely, which made discussions with him very inspiring. Prof. Giel Paardekooper encouraged me by his enthusiasm whenever my study took a turn which he recognized to be interesting. He always animated me by his broad knowledge of mathematics and wide interests. Not long after starting my job at Tilburg University I learnt a lot from Dr. Jacob Engwerda when he made me co-author of an article for a scientific journal. Dr. Jacob Engwerda continued to take an enormous effort in careful and critical reading of any text I wrote giving numerous valuable comments. He helped me in all possible ways. My three advisors did much more. I appreciate their friendly moral support and personal interest during the past four years. Furthermore, they also organized the visits of important researchers in the field of identification. These researchers, Prof. Manfred Deistler, Prof. Peter Young and Prof. Rainer Dahlhaus, have all influenced my work quite a lot. I thank these researchers for their interest and sympathy. By the initiative of my three advisors we also participated in a Science project of EEC, which enabled me to stay for a month in the Centre for Research on Environmental Systems and Statistics of Prof. Peter Young in Lancaster, England. It was a very fruitful visit, where I had good discussions not only with Prof. Peter Young, but also with Dr. Wlodek Tych. I very much enjoyed the hospitality and friendship that he, his wife and daughter, and also Prof. Young and his wife, offered me.
Many others have contributed to the succes of this work, like Dr. Bernard Hanzon who introduced me to the theory of Schur parameters and Detlef de Vries who collaborated with me on a specific subject. My room mate Anne and the other collegues of Tilburg University were always ready to help me and created a nice working atmosphere. I thank them all.
This study would also not have been possible without the education I received at the Mathematical Institute of the Catholic University of Nijmegen. I wish to express my thanks to the teachers there, especially to Prof. Van Rooy.
Finally I have to mention the tremendous moral and practical support I received from my mother and father, my brothers, sister and other relatives. Despite his illness, my father gave all his energy to motivate me and to assist me in typography, lay-out and language matters, helped by my brother André. To my parents, I dedicate this work.

Arie Rijkeboer
March 1994

I wish to thank Professor Peter Young of Lancaster University, England, Professor Rainer Dahlhaus of Universität Heidelberg, Germany, Dr. Bernard Hanzon of Vrije Universiteit Amsterdam, Professor Ben van der Genugten, Dr. Jacob Engwerda, Professor Hans Schumacher en Professor Giel Paardekooper of Tilburg University for their willingness to participate in the Ph.D. committee and for their comments.

Arie Rijkeboer
March 1994

Contents

Chapter 1

Introduction

1.1 General motivation

This thesis is concerned with non-stationary time series—that is with time series in which the statistical pattern of fluctuations is itself subject to change over time. Such data series may arise for instance from systems that are influenced both by short-term random disturbances as well as by longer-term changes of operating conditions. Examples of such situations are manifold. Consider for instance ecological data, say tree ring measurements. The widths of the rings found in a tree stem reflect certain factors exhibiting a natural variation, such as the amount of rainfall in a given year, but they may also reflect changes in environmental conditions, such as the acidity of the soil. In industrial data, the performance of a machine over time may be influenced by several short-term random factors, but will also be influenced by wear and tear; there could also be sudden changes of pattern caused by a failure of some part of the machine. In medicine, data such as obtained by successive measurements of, e.g. the blood sugar level of some patient, will show some random fluctuations perhaps related to the patient's eating behavior, but changes in the pattern of the fluctuations may give indications about the progress of a disease or the success of treatment. Economical data usually exhibit considerable randomness as well as short-term correlation, but again changes in pattern may be related to structural changes in the economy.

The cited examples suggest that, in the study of non-stationary time series, usually the main focus of interest will be the change of operating conditions rather than the short-term correlations. Nevertheless, one is forced to model the short-term pattern if one is to give a precise analysis of the *change* of the pattern. The key issue is then, how to build a model for the nonstationarity that can be effectively linked to models for current patterns. This issue will be addressed in the chapters to follow by the construction of a certain class of models for non-stationary time series. The construction is based on a description of 'current patterns' by means of standard autoregressive equations of a fixed order. The 'operating conditions' are thought of as determining a current pattern, so a particular autoregressive equation. The change of operating conditions in this framework then becomes a motion in the set of autoregressive models of a fixed order. To describe this motion one needs a notion of distance, and for that purpose the set of autoregressive models of a fixed order is equipped with the structure of a Riemannian manifold. The

choice of the Riemannian structure is a modelling decision, which ideally should be based on an assessment of possible magnitudes of variations of operating conditions, and the relation between current operating conditions and the current autoregressive equation.

The class of models that will be considered in this thesis may be viewed as a class of autoregressive models with time-varying parameters, together with a description of the variation of parameters that at first sight may seem rather complicated. In the literature (cf. for instance Ljung and Gunnarson, 1990) similar models have been proposed which, however, used simpler dynamics for the parameter variation. In the terminology of this thesis, such models may be viewed as resulting from the choice of a particular Riemannian structure (the 'Euclidean' one). By making this choice, the modeller assumes that differences of coordinate values are a good measure of changes in operating conditions. This assumption is justifiable by a linearization argument if the operating conditions remain close to some nominal conditions. However, for large changes this argument is no longer valid, and it may be useful to carry out re-parametrizations in response to changes of the operating point. Such a re-parametrization is provided automatically by the choice of a Riemannian structure.

This is the method of modelling that is the subject of the present thesis. The work includes the construction of the mathematical framework, a discussion of considerations that may play a role in the choice of a particular Riemannian structure, and an investigation of stability properties and of the notion of 'current pattern'. Although the emphasis is on modelling, a discussion of the identification problem is also included; a complete statistical analysis of the proposed model, however, is outside the scope of the thesis. It will be assumed that the analysis of data takes place off-line so that computationally demanding procedures are not immediately ruled out. Computational complexity will be kept as a consideration in a global sense, without going into the details of optimizing numerical speed and reliability.

1.2 Previous research

The statistical treatment of *autoregressive processes with constant coefficients* goes back to the work of Mann and Wald (1943), who were motivated by economic applications. From their paper we cite:

> The values of a great many variables important in the study of economics depend on the values previously assumed by these variables. For example present prices of a certain set of commodities may depend on the prices of these commodities in previous time periods, etc. Such a relationship is usually described by a set of stochastic difference equations.

These equations include the autoregressive form

$$y_t - a_1 y_{t-1} + \cdots - a_n y_{t-n} = \epsilon_t \tag{1.1}$$

where y_t are the observations, the $(\epsilon_t)_{t\in\mathbf{Z}}$ are i.i.d. with mean zero and finite higher moments, and the zeros of the polynomial $x^n - a_1 x^{n-1} + \cdots - a_n$ are smaller than one in absolute value. Mann and Wald give expressions for the maximum likelihood estimates

$\hat{a}_T$ of the coefficients $a = (a_1, \cdots, a_n)$ based on T observations and prove consistency and asymptotic efficiency of these estimates.

The autoregressive process $(y_t)_{t \in \mathbf{Z}}$ defined by equation (1.1) is an example of a *stationary* time series, which is to say that the joint probability distribution of k subsequent elements $y_{t-k}, \cdots, y_{t-1}$ of the process does not depend on t, for all $k \in \mathbf{N}$. The word 'stationary' was introduced and the systematic study of stationary time series began with an article of Khinchin in 1934 (see the historical references in Rozanov (1967)). The theory of this type of time series has been developed to a great extent and is widely applicable. However, the assumption of stationarity is not always justified. Box and Jenkins (1970) state in their preface to *Time Series Analysis, Forecasting and Control*:

> In particular, time series are often best represented by *non-stationary* models in which trends and other pseudo-systematic characteristics which can change with time are treated as statistical rather than as deterministic phenomena. Furthermore, economic and business time series often possess marked seasonal or periodic components themselves capable of change and needing (possibly non-stationary) seasonal statistical models for their description.

Their book treats the famous non-stationary ARIMA (autoregressive integrated moving-average) models in which stationarity is obtained only after differencing the given time series a number of times.

Although ARIMA models are capable of representing adequately some types of non-stationarity, they are much too limited to serve as a general model class for non-stationary series. This thesis will deal with a different model class of considerably larger descriptive power: (univariate) *autoregressive processes with time-varying coefficients.* Models of this type have already been used in the literature for a wide variety of phenomena, ranging from earthquakes (Kitagawa and Gersch,1985) to electro-encephalograms (Jansen, 1979). Other examples include speech, vibration in machinery (Gersch, 1982), and flutter in airplanes. In a somewhat different context, models for time-varying autoregressive processes with exogenous variables (time-varying ARX processes) play an important role in on-line identification for adaptive control. A survey of methods used in this area can be found in Ljung and Gunnarsson(1990).

In the field of research on time-varying autoregressive processes and, more generally, varying-parameter models, two important lines of development can be distinguished which differ in their modelling of the variation of parameters. First of all, the coefficients may be modeled as *deterministic* functions of time. This approach is usually linked to representations in the *frequency domain.* Remember that, for a (univariate) stationary process, the correlation of a process element at time t with a process element at time $t-k$ only depends on the time difference k. A (univariate) stationary process is characterized by one, time-invariant spectral measure that determines these correlations. In view of this characterization, one may attempt to assign time-varying spectral measures to certain classes of non-stationary processes. Priestley used the term 'semi-stationary' for such a class of non-stationary processes, and 'evolutionary spectral density' for the time-varying spectral densities that he defined for these processes Priestley (1965, 1966, 1981). Unfortunately, Priestley's definition suffers from a lack of uniqueness: various evolutionary spectral densities can be assigned to the same process, depending on the

white noise process chosen as reference point.

Another approach using deterministic coefficient models and frequency-domain descriptions is due to Dahlhaus (1993). In this approach, a non-stationary process is not viewed in isolation, but rather as a member of a family. Dahlhaus defined a specific class of such families of processes that he called 'locally stationary'. He succeeded in assigning a uniquely defined time-varying spectral density to the families of this class. It is possible to give simple conditions for a sequence of time-varying autoregressive processes to be locally stationary. The time-varying spectral density is the basis for the procedures given by Dahlhaus for the fitting of parametric, deterministic time-varying autoregressive coefficient models to a set of observations, considered to be the realizations of a process in a locally stationary family. Dahlhaus developed a complete statistical theory for these procedures.

A second direction of research uses *stochastic* models for parameter changes, often in combination with *time-domain* descriptions. To give an example of such a class of models, the coefficients $\vec{a}_t = (a_{1t}, \cdots, a_{nt})^T$ of the time-varying autoregressive process

$$y_t = a_{1t}y_{t-1} + \cdots + a_{nt}y_{t-n} + \sigma_t\epsilon_t, \quad \epsilon_s \overset{i.i.d.}{=} \mathcal{N}(0,1)\ \forall s \tag{1.2}$$

can be modelled for instance as a random walk:

$$\vec{a}_{t+1} = \vec{a}_t + \tau_t\vec{\lambda}_t \quad \vec{\lambda}_s \overset{i.i.d.}{=} \mathcal{N}(0, I_n)\ \forall s, \quad \vec{\lambda}_s \text{ independent of } \epsilon_t\ \forall s,t. \tag{1.3}$$

The coefficients can then recursively be estimated by Kalman filtering (and smoothing), optimally in the sense that the conditional estimation error variance given the data is minimized. The Kalman filter allows estimation of the states κ_t from the observations y_t in the linear dynamic regression model or system consisting of the transition equation

$$\kappa_{t+1} = A_t\kappa_t + g_t\lambda_t \tag{1.4}$$

and the measurement equation

$$y_t = x_t'\kappa_t + \sigma_t\epsilon_t \tag{1.5}$$

in which $(\lambda_t\ \epsilon_t)_{t\in\mathbf{Z}_+}^T$ is unit variance Gaussian white noise. The x_t's are here considered as regressor vectors that in case of a model for a time-varying autoregressive process of order n depend linearly on the preceding observations $y_{t-1}, \cdots, y_{t-n}$. The filter construction requires the specification of the variance factor g_t and the matrices A_t in the transition equation and the noise level σ_t^2 in the measurement equation. Also the initial state estimate $\kappa_{0|-1}$ with its error variance $\Sigma_{0|-1}$ has to be specified (or, equivalently, a prior Gaussian distribution for κ_0). The entries of the matrices $g_t, A_t, \Sigma_{0|-1}$, the vector $\kappa_{0|-1}$ and the numbers σ_t^2 are parameters that determine the model and are sometimes called hyper(-structural) parameters (Harvey, 1990, Kitatagawa and Gersch, 1985).

As for instance pointed out by Kitagawa and Gersch (1985), Kalman smoothing for the stochastic coefficient model (1.2–1.3) given the observations $y_{-n}, \cdots, y_{-1}, y_0, \cdots, y_T$ has a very natural interpretation because it is just deterministic smoothing based on the minimization of the criterion

$$\begin{aligned} J_T \;=\; & \sum_{t=0}^{T} \frac{(y_t - a_{1t}y_{t-1} + \cdots - a_{nt}y_{t-n})^2}{\sigma_t^2} + \\ & + \sum_{t=0}^{T-1} \frac{||\vec{a}_{t+1} - \vec{a}_t||^2}{\tau_t^2} + \langle \vec{a}_0 - \vec{a}_{0|-1}, \Sigma_{0|-1}^{-1}(\vec{a}_0 - \vec{a}_{0|-1})\rangle. \end{aligned} \tag{1.6}$$

Kitagawa and Gersch mention that criteria of this type date back to the the early twenties, notably the work of Whittaker (1923). The first term is a sum of squared 'prediction errors'. In a regression problem of the type $y_t = x_t'\kappa + \sigma_t\epsilon_t$, $(\epsilon_t)_{t\in\mathbf{Z}}$ white noise, with *time-invariant* parameter vector κ, a least squares estimate of κ would only require the minimization of the sum of squared prediction errors $\sum_{t=0}^{T}\sigma_t^{-2}(y_t - x_t'\kappa)^2$. In the *time-varying* case, the two other terms in the criterion (1.6) are needed, to guarantee the existence of a unique solution $\vec{a}_{0|T}, \cdots, \vec{a}_{T|T}$. These terms can be interpreted as penalties for changing the coefficients. The hyperparameters τ_t^2, σ_t^2 and the entries of $\Sigma_{0|-1}$ are weights in this criterion. The virtue of the model (1.2–1.3) in comparison with just the deterministic criterion (1.6) is that it gives a guideline how to determine these weights: as a stochastic model it suggests Maximum Likelihood Estimation. Although this way of estimating the hyperparameters from the data is not always easy, the power of a stochastic coefficient model (in comparison with a parametric deterministic coefficient model) is that a relatively small number of (hyper-)parameters is sufficient in order to obtain a good fit.

Stochastic coefficient modelling does not exclude work in the frequency domain. This is certainly true for the determination of hyperparameters. Spectral techniques were used by Young, Ng, and Armitage (1989, see also Young, Ng, Lane, and Parker, 1991) to estimate the ratio between transition noise variance and measurement noise variance (NVR) in non-stationary time-varying stochastic parameter models like the random-walk-plus-noise or trend model and models with trend and seasonal components. They present this as a first step for ML estimation methods and show that the spectral techniques are also useful for decomposition of the time series in trend and seasonal components.

Kitagawa and Gersch (1985), see also Kitagawa (1983), have introduced the idea of an 'instantaneous spectral density' for a time-varying process which they used in the context of coefficients following an (integrated) random walk. They estimated the instantaneous spectral density simply by means of the coefficient estimates obtained from Kalman filtering and smoothing:

$$f_t(\phi) = \frac{\sigma_t^2}{|e^{in\phi} - a_{1t}e^{i(n-1)\phi} + \cdots - a_{nt}|^2}. \quad (\phi \in [0, 2\pi]) \tag{1.7}$$

One may view this as an extension of the estimation of the spectral density for stationary processes by fitting time-invariant autoregressive models, as explored for instance by Berk (1975) . Dahlhaus (1993) has given a rigorous meaning to the idea of Kitagawa and Gersch of an instantaneous spectral density by connecting it to certain sequences of time-varying autoregressive processes with *deterministic* coefficients. The same will be done in this thesis for certain sequences of time-varying autoregressive processes whose coefficients follow a *smoothly integrated random walk* (see chapter 4). The obtained time-varying spectral density will then be conditional on the stochastic process generating these coefficients. This concept of a time-varying spectral density for a sequence of processes will be linked to a concept of *stability* for such a sequence in almost the same way as in the time-invariant, single process case. Stability too will be conditional on the stochastic process generating the coefficients. Although it will not be shown in this thesis, we think that these concepts open the way for fitting procedures based on spectral techniques (and their statistical analysis) for hyperparameters of models with stochastic

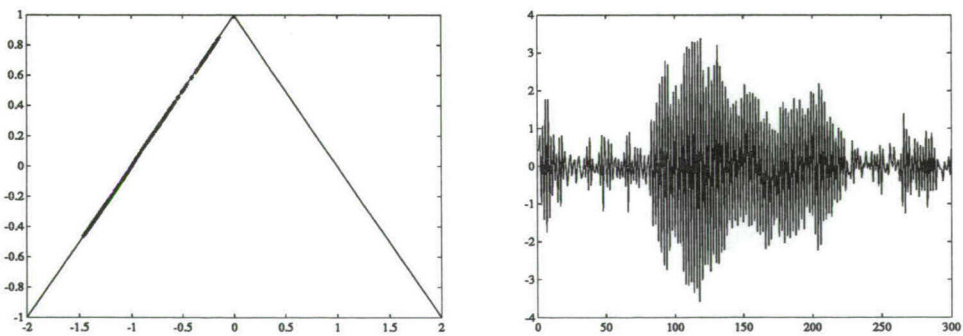
coefficients, especially hyperparameters that affect the measurement equation: those on which the regressors x_t depend in more complicated models.

1.3 Proposed approach

As already mentioned above, an attractive feature of stochastic coefficient models for time-varying autoregressive processes is their parsimony in (hyper-)parameters in comparison with parametric deterministic coefficient models. Economy in parameters is a natural requirement for models of time series. This is reflected for instance in Akaike's Information Criterion (AIC) (Akaike, 1974) that is widely used for selection of a model on the basis of a finite number of data from a set of models with different numbers of parameters. The criterion contains a term that sets a penalty on the number of parameters used.

One aim in this thesis is to *refine* the random-walk coefficient model in order to obtain better fits without sacrificing parsimony in hyperparameters. Other aims will be to obtain models in which hyperparameters have a clear interpretation that could facilitate their determination, and to obtain models with guaranteed stability properties. The refinement lies in the choice of the transition noise variance.

Perhaps the best way to introduce this idea is by means of a simple example. As is well-known, an AR(2) model with constant coefficients generates a stationary process if and only if its coefficient vector $\vec{a} = (a_1, a_2)$ lies in the open triangle with corner points $(-2, -1)$, $(2, -1)$ and $(0, 1)$. The time series in Fig.1.1 was generated as a time-varying AR(2) process with constant measurement noise and coefficients lying near the upper edge of the stability triangle.



time-varying coefficients generated series

Figure 1.1:Time-varying AR(2) process

The time series looks stable although not stationary. In order to estimate the coefficients, let us apply a parsimonious model with stochastic coefficients. We use a time-varying AR(2) model with constant measurement noise level, where, as usual, the coefficients are modelled as a random walk. For the sake of parsimony, the transition variance is supposed to be a multiple of the identity and constant in time. Hence, we use just the

model (1.2–1.3). According to the maximum-likelihood criterion, one can estimate the measurement noise level and the transition variance, and by the Kalman filter one can obtain the filtered estimates of the coefficients. The curve of these filtered estimates of coefficients is given in Fig.1.2.

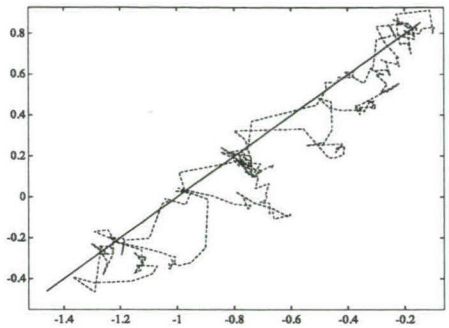


Figure 1.2:Estimates of coefficients

It is a rather ambiguous curve; many estimated coefficients lie outside the stability triangle. This suggests that a better fit of the model to the data can be obtained by impeding the motion of the coefficients in the direction perpendicular to the edge of the triangle. This can be done by changing the transition variance matrix from a multiple of the identity to a positive definite matrix with two eigenvectors, the one corresponding to the largest eigenvalue parallel to the edge of the triangle, and the other one perpendicular to this edge. In this way the parsimony of the model is decreased and the complexity increased: we have to estimate two eigenvalues of the transition variance matrix instead of one. Apart from this, the method is very *ad hoc*. If, in another case, generated coefficients would follow a non-straight curve, the identifier would be tempted to use different transition variance matrices at different times in order to get a better fit. This would increase the number of hyperparameters even further. Can there be an alternative where it is not necessary to sacrifice parsimony?

We think that there sometimes is. The intuitive idea is to make the transition variance a function of the current value of the coefficients. We replace the transition equation

$$\vec{a}_{t+1} = \vec{a}_t + \tau\vec{\lambda}_t \tag{1.8}$$

by:

$$\vec{a}_{t+1} = \vec{a}_t + \tau g(\vec{a}_t)\vec{\lambda}_t. \tag{1.9}$$

Here, g is a square matrix function. Instead of a constant transition variance $\tau^2 I_n$, we get a transition variance $\tau^2 g(\vec{a}_t)g'(\vec{a}_t)$ that is a function of the most recent value of the coefficient vector. As a result, the coefficient process is not a random walk in the strict sense anymore, but a *martingale* in general. In the transition equation (1.9) different square matrix functions g with the same product function gg' yield coefficient processes with the same probability distribution. One can conceive of such a square

matrix function function g or, equivalently, gg' as an *a priori* view on which direction the coefficient process is most likely to move from a certain position. For instance, in a time-varying AR(2) model one can use the function gg' to 'penalize' the direction perpendicular to any edge of the stability triangle, whenever the coefficient process comes near the edge.

We do not suggest to estimate such a function gg' from the data. (At most, in models with one transition-equation-hyperparameter τ, one can estimate an optimal convex combination of a given finite set of transition variance functions gg' by maximum likelihood procedures.) The choice of such a function should be considered as a modelling decision, comparable for instance to deciding upon a linear model in the logarithms of the original variables. Considerations that play a role in such decisions include a feeling for the data, and for the process that generated the data; they may also include arguments of convenience, such as ease of computation in optimal estimation, stability properties, and interpretability of the hyperparameters.

Adaptive modelling of the transition variances and the measurement noises is a popular tool in on-line identification schemes for adaptive control. Since transition variances and measurement noises are key parameters in the equations of the Kalman filter, the identifier with adaptive control purposes has to make good choices for these parameters on-line. A standard technique for this is Recursive Least Squares. In the RLS method, at every instant t a weighted sum of squared prediction errors $(y_k - x_k'\kappa)^2$ until time t is minimized in which the most recent prediction errors have much larger weights than the older ones; therefore, the latter are gradually forgotten. In this way one takes account of the changing underlying structure of the process that is currently identified and one obtains a time-varying sequence of estimated coefficients $\hat{\kappa}_0, \cdots, \hat{\kappa}_t, \cdots$, which does not necessarily converge. It has been shown by Ljung and Gunnarsson (1990) that this technique can be understood as an application of a Kalman filter with transition and measurement variances specially chosen (or 'estimated') as functions of the past observations. These 'estimates' do not carry a direct relation to the current value of the coefficients and hence the method is of a different kind than the methods discussed in this thesis. We also note that we will emphasize off-line identification in which smoothing (i.e. updating past coefficients to take account of observations coming in later) plays an important part.

The intuitive idea of the transition variance as a positive definite matrix function of the current coefficients actually needs some refinement. Here, differential geometry enters our considerations. A positive definite smooth matrix function turns its domain into a Riemannian manifold: the inverse of the given function defines a *distance* on the coefficient space. In this way, the choice of the matrix function can be seen as an *a priori* view of the distances that the stochastic coefficient process is faced with.

1.4 Outline of the thesis

This thesis consists of seven chapters, of which the first is the current introduction. The second chapter will provide preliminary material concerning stochastic processes and differential geometry; nearly all of this material is known and is collected here

for the convenience of the reader. The actual development starts in chapter 3. Here we generalize the idea of a random walk to that of a a (special) 'polygonal process of independent directional variation on a Riemannian manifold'. In order to establish the links with the local stationarity concept, referred to above, we also generalize the idea of a smoothly integrated random walk to that of a general polygonal process of independent directional variation on a Riemannian manifold. We shall show that there exist manifolds that are bounded subsets of $\mathbb{R}^n$, and polygonal processes remaining on these for all time *without* converging. At the end of the chapter, we give some very simple statements on convergence for this type of processes. These statements are also valid for random walks that actually are just a special form of polygonal processes. On the other hand, it will depend on the properties of the positive definite matrix function, defining the Riemannian manifold, whether polygonal processes on it can have infinite lifetime. A stopping time (called *running time*) will be introduced for the lifetime of a polygonal process.

We shall give precise conditions under which it is possible to define other coordinates, than the usual ones on the coefficient space, such that the coefficient process expressed in these coordinates is just a martingale or even an ordinary (smoothly integrated) random walk. In the latter case, using these coordinates, the model will have a *linear* transition equation (and a non-linear measurement equation, because the coefficients are possibly non-linear functions of these coordinates). This is preferable to a model with a nonlinear transition equation.

We shall formulate an important class of models, called the Geodesic Models, where the coefficients follow a polygonal process, but with (scalar) linear transition equation on some coordinates. It will be shown that one can obtain such models for any Riemannian manifold. The number of hyperparameters for a Geodesic Model for a time-varying AR(n) process is $2n + 2$, or $2n + 4$ if the polygonal process is a smoothly integrated random walk on some coordinates, whereas for a usual model with transition equation as in (1.8) the number of hyperparameters is $n + 3$ (not counting initial data in both cases). For small values of n, it will be worth considering the use of a Geodesic Model on a suitable Riemannian manifold in order to obtain a better (and, in some way, smoother) fit instead of a simple increase in the order of the usual model.

Up to this point, we only considered autoregressive processes $(y_t)_{t\in\mathbf{Z}}$, $y_t = a_{1t}y_{t-1} + \cdots + a_{nt}y_{t-n} + \sigma\epsilon_t$, where $(\epsilon_t)_{t\in\mathbf{Z}}$ is Gaussian unit variance white noise and only the coefficients $\vec{a}_t = (a_{1t}, \cdots, a_{nt})$ change with time. Throughout the thesis we shall also consider variation in time of the noise level σ_t^2. We shall call the noise level a "coefficient" as well and model the time variation of the noise level jointly with the time variation of the other coefficients. Instead of the stability region $AR_n \subset \mathbb{R}^n$, i.e. the set of coefficients $\vec{a}$ corresponding to time-invariant stationary AR(n) processes with measurement noise level 1, we consider the stability region $U_n \subset \mathbb{R}^{n+1}$, which is defined as the set of coefficients $(\vec{a}, \sigma^2)$ corresponding to time-invariant stationary AR(n) processes with arbitrary noise levels. Geometrically, U_n is just a cone constructed from AR_n.

Intuitively speaking, there must be a concept of stability for time-varying AR processes just as for time-invariant ones. A definition of a concept of stability for time-varying processes will be given in chapter 4. For a time-invariant AR(n) process it is necessary for stability that the coefficient vector belongs to a certain bounded and con-

nected open subset AR_n of $\mathbb{R}^n$ (the triangle in the AR(2) case). We shall show that for time-varying AR(n) processes our concept of stability requires that coefficients mainly lie in the same stability region AR_n. The theory of Riemannian manifolds will be very helpful in constructing models with good stability and linearity properties. Guaranteed stability properties will be related to completeness of the Riemannian manifold structure on AR_n, defined by the transition variance function. We shall see that it will still be possible to perform asymptotic analysis (for instance of statistical procedures) even if we work with time-varying AR processes of finite lifetime. This will be done by following Dahlhaus' treatment of asymptotics, as expressed in his concept of local stationarity, which is closely related to the concept of stability.

In chapter 5 we discuss three particularly interesting Riemannian structures on the stability region AR_n. The Riemannian structures, introduced here, include the Euclidean structure of the stability region as a submanifold of $\mathbb{R}^n$. With no restriction on the coefficient space, the Euclidean structure yields the usual model. If the stability region is equipped with this structure, it becomes a Riemanian manifold that is geodesically incomplete. As a consequence, the assumption of stability for all time (in the way we define it) cannot be incorporated in models with constant hyperparameters based on the Euclidean structure. But we shall also give a Riemannian manifold structure which is complete and, hence, gives rise to models with guaranteed stability properties, while the coefficient process does not converge. The structure we shall give will at the same time give rise to models which nearly always have a linear transition equation on suitable coordinates.

The third Riemannian structure we shall discuss will be the one induced by the asymptotic Fisher information matrix for time-invariant stationary AR processes. The geometry of the set of stationary AR models, based on this Riemannian structure, is very important for the statistical analysis of such AR processes as shown for instance by Amari (1987). In a model based on this structure, the hyperparameter τ^2 (or, time-varying, τ_t^2) in transition equation (1.9) gets a clear interpretation, not depending on the actual values of the coefficients.[1] Indeed, in the corresponding Whittaker criterion

[1]Let us give a small example of what we mean by interpretability of the hyperparameter. Take the following simple AR(1) model:

$$\begin{aligned} a_{t+1} &= a_t + \tau_t g(a_t)\lambda_t \\ y_t &= a_t y_{t-1} + \epsilon_t \end{aligned} \tag{1.10}$$

where $(\lambda_t)_{t\in\mathbb{Z}}$ and $(\epsilon_t)_{t\in\mathbb{Z}}$ are independent sequences of white noise; $\lambda_t \stackrel{d}{=} \mathcal{N}(0,1)$ and $\epsilon_t \stackrel{d}{=} \mathcal{N}(0,1)$ for all $t \in \mathbb{Z}$. Suppose that $\tau_t = 0$ if $t \neq k$ and $\tau_k = \tau$. Then the system jumps at time $t = k$ from one AR(1) system with constant coefficient to another. If we take $g = 1$, then the jump from coefficient $a_t = 0.1$ $(t \leq k)$ to coefficient $a_t = 0.05$ $(t > k)$ receives the same probability from the model as the jump from coefficient $a_t = 0.9$ $(t \leq k)$ to $a_t = 0.95$ $(t > k)$. Yet, the Kullback-Leibler divergence between AR(1) systems before and after the first jump is far smaller than the divergence between the systems before and after the second jump (see D. de Vries 1992). In some situations, one may prefer to have a model where τ_k^2 can be interpreted as the mean Kullback-Leibler divergence between systems before and after the jump, independently from where the coefficients are in the parameter space. Note that for small τ_k^2, the Kullback-Leibler divergence and the squared distance of the asymptotic Fisher metric are almost equal:

$$KL(a_{t+1}, a_t) \tilde{\sim} (a_{t+1} - a_t) I(a_t)(a_{t+1} - a_t) = \tau_k^2 g(a_t) I(a_t) g(a_t) \lambda_t^2. \tag{1.11}$$

the weight τ_t signifies the ratio between the penalty to a change in coefficient at time t and the expected contribution of observation y_t to the information about the coefficients. It will be shown that in a Geodesic Model based on the asymptotic Fisher metric, assuming a scalar transition equation with constant hyperparameter τ, asymptotically each observation contributes in expectation the same amount of information about the changing coefficient coordinates.

The three Riemannian structures will be extended to the stability region U_n. The third structure, referred to above, will be extended to the one defined by the asymptotic Fisher information matrix with respect to $(\vec{a}, \sigma^2)$ of the time-invariant stationary AR(n) process with coefficients $(\vec{a}, \sigma^2)$, and the other structures will be extended in the same way. It is shown that these structures then satisfy a criterion, such that not only the Geodesic Models have a linear transition equation on some coordinates, but also a class of more flexible models; the Geodesic plus Noise Models. Later on it will be shown that in the latter models the Gauss-Newton method for posterior mode estimation can be split in two parts; one for updating the noise levels, and the other for updating the coordinates of the other coefficients.

Chapter 6 will be concerned with identification aspects of the proposed modelling framework. Already by replacing the usual coefficient random walk by a coefficient polygonal process, we introduced non-linearities in the 'linear dynamic regression model'. Identification by the ordinary Kalman filter plus smoothing will not be possible anymore. Instead one may use an approximation, the Extended Kalman Filter and its smoothing version. The properties of the Extended Kalman Filter are much more difficult to establish than those of the standard filter, see for instance Anderson and Moore (1979); Jazwinski (1970). In case the model still has a linear transition equation, we shall show (generalizing a result of Fahrmeir and Kaufmann) that we can understand the Extended Kalman Filter and its smoothing version as a simplification of a Gauss-Newton method in posterior mode estimation, which, if one wishes, could also be carried out in full. In case of a non-linear transition equation, this relation is not clear at all.

The estimation problem of the hyperparameters in Geodesic Model, Geodesic plus Noise Model and a simple model with transition equation equal to (1.9) (in which usually there are no coordinates, such that the transition equation expressed in these is linear) will also be addressed in chapter 6. The proposed estimation procedures are mostly quite standard (we follow mainly procedures also utilized by Harvey (1990), Kitagawa and Gersch (1985), and a proposal of Fahrmeir and Kaufmann (1991); however, a new statistical analysis is needed because of the particular purpose for which we use these procedures. A full analysis of this kind will not be undertaken in this thesis. We merely illustrate that the proposed methods in case of Geodesic Model and the 'simple' model work reasonably well in a number of experiments with simulated data.

Chapter 7 concludes the thesis. Here we summarize the conclusions of the work and present recommendations for further research. In particular, we list a number of interesting open problems.

---

where $I(a_t) = \frac{1}{1-a_t^2}$ is the asymptotic Fisher information in the coefficient a_t. So, the conditional expectation of $KL(a_{t+1}, a_t)$ given a_t will be approximately equal to τ_k^2, if we choose $g(a_t) = \sqrt{1-a_t^2}$.

Chapter 2

Preliminaries

In this chapter, we give some basic concepts and results from several theories that will be used frequently. The first section gives a brief introduction of differential geometry. The second section elaborates some concepts in the first section a little bit further. The concepts in the third section are not well known, as far as we know. In this section, we introduce the notions of conditional expectation and variance of a stochastic variable with values in the tangent bundle of a Riemannian manifold. The fourth section contains the well known expression for the asymptotic Fisher information matrix w.r.t. coefficients and noise level for the set $\mathcal{U}_n$ of stationary time-invariant AR(n) processes. In the fifth section, we give a brief summary of a well known theory by which important parametrizations of this set $\mathcal{U}_n$ can be constructed: the theory of shift invariant inner products. The Levinson algorithm of fitting an AR(n) model to a set of univariate observations by minimizing a sum of squared prediction errors can be based on this theory. In the sixth section, we give some references on the most important identification algorithm for linear AR(n) models with stochastic coefficients: Kalman Filtering and Smoothing. We also refer to some literature on its extension, Extended Kalman Filtering and Smoothing, designed to estimate states in a non-linear dynamical stochastic model. We come back to these algorithms in chapter six extensively, hence we skip an introduction of them here. The final section of this chapter gives an important well known elementary result of the theory of compact metric spaces: the theorem of Arzela-Ascoli. As we have said in the introduction, we are mainly interested in autoregressive processes with *slowly* varying coefficients. In chapter four, we shall formalize this idea to the concept of *equislowly* varying coefficients and base Dahlhaus's theory of asymptotics on it. The theorem of Arzela-Ascoli is essential in this construction. In this section, we also recapitulate some aspects of the Ornstein-Uhlenbeck Velocity process. These aspects will also be used in constructing the base for an asymptotic theory.
The first three sections are preparations for chapter three, the fourth and fifth section for chapter five, the sixth section for chapter six, and the seventh section for chapter four.

2.1 Basic ideas for Riemannian geometry

References, used on differential geometry, are Helgason (1978), Kobayashi and Nomizu (1963, 1969). In the book of Amari (1990) a very good introduction to differential geometry can be found. In that book a direct relation with statistical methods is given. Differential geometry is also used in non-linear dynamical control systems. In the introduction of the book of Nijmeijer and Van der Schaft (1990) other aspects of differential geometry are given than in the book of Amari. Both introductions are relevant.
This section aims to give a short exposition of main ideas in differential geometry. In this introduction, we restrict ourselves to an open subset M of $\mathbb{R}^n$, for example $M = AR_n$, the stability region for coefficients of time-invariant autoregressive processes with measurement noise level 1. The section, following the present one, will be more general.
We can consider M as an n-dimensional manifold. Then, in any point θ of M, we can define an n-dimensional linear tangent space, consisting of all possible tangent vectors $\frac{\partial \gamma}{\partial t}_{|0}$ of differentiable curves $t \mapsto \gamma(t)$ on M, such that $\gamma(0) = \theta$.
Any C^∞-differentiable map $\chi : V \to M; (v_1, \ldots, v_n) \mapsto \chi(v_1, \ldots, v_n)$ with C^∞-differentiable inverse $\chi^{-1} : M \to V$ is called a *parametrization* of M. The corresponding tangent vectors $\frac{\partial \chi}{\partial v_i}_{|\hat{v}}$ in the point $p = \chi(\hat{v}) \in M$ will be denoted by $\frac{\partial}{\partial v_i}$, and form a basis of the tangent space in $p = \chi(\hat{v})$. The tangent space in P itself will be denoted by M_p.
If $\psi : \tilde{V} \to M; (w_1, \ldots, w_n) \mapsto \psi(w_1, \ldots, w_n)$ is another parametrization of M, then $\chi^{-1} \circ \psi : \tilde{V} \to V; (w_1, \ldots, w_n) \mapsto (v_1, \ldots, v_n)$ is the coordinate transformation with derivative $D\chi^{-1} \circ \psi$, which transforms the tangent space in $\chi(v) = \psi(w)$ in the following way:

$$\frac{\partial}{\partial w_j} = \frac{\partial \psi}{\partial w_j} = \frac{\partial \chi(\chi^{-1} \circ \psi)}{\partial w_j} = \sum_{i=1}^{n} \frac{\partial \chi}{\partial v_i} \left(\frac{\partial (\chi^{-1} \circ \psi)_i}{\partial w_j} \right) = \sum_{i=1}^{n} (D\chi^{-1} \circ \psi)_{ij} \frac{\partial}{\partial v_i}. \tag{2.1}$$

Note that the trivial parametrization $id : M \to M; (b_1, \ldots, b_n) \mapsto (b_1, \ldots, b_n)$ gives the tangent vectors $\frac{\partial}{\partial b_1} = (1, 0, \ldots, 0)', \ldots, \frac{\partial}{\partial b_n} = (0, 0, \ldots, 1)'$.
A *vector field* X on M is a map from M to the union of all the tangent spaces M_p, $p \in M$, such that $X(p) \in M_p$ for all $p \in M$. The vector field X is called C^∞, if for some parametrization $\chi : V \to M; (v_1, \ldots, v_n) \mapsto \chi(v_1, \ldots, v_n)$ we have:

$$X(p) = \sum_{i=1}^{n} f_i(v) \frac{\partial}{\partial v_i}(p), \tag{2.2}$$

where the functions $f_i : V \to \mathbb{R}$ are C^∞. With equation (2.1) it is clear that then also $X(p) = \sum_{j=1}^{n} g_j(w) \frac{\partial}{\partial w_j}(p)$, where $g_j : \tilde{V} \to \mathbb{R}$ is C^∞.
Now, suppose that we have defined an inner product $\langle , \rangle_p$ in the tangent space M_p in any point $p \in M$.
If there is a parametrization $\chi : V \to M$, such that the matrix function $H_\chi : V \to \mathbb{R}^{n \times n}$;

$$H_\chi(v) = \left(\langle \frac{\partial}{\partial v_i}, \frac{\partial}{\partial v_j} \rangle_{|\chi(v)} \right)_{ij} \tag{2.3}$$

is C^∞-differentiable, then for any parametrization ψ the matrix function H_ψ will be C^∞-differentiable, because from the tangent space transformation above it follows that:

$$H_\psi(w) = D'\chi^{-1} \circ \psi_{|w} H_\chi(\chi^{-1} \circ \psi(w)) D\chi^{-1} \circ \psi_{|w}. \tag{2.4}$$

In this case, we shall call M a Riemannian manifold and $H_\chi(v)$ the metric on M w.r.t the parameters $(v_1, \ldots, v_n)$. In order to define a Riemannian manifold structure on M, it suffices to give one parametrization $\chi : V \to M$ and a C^∞-differentiable symmetric matrix function $H_\chi : V \to \mathbb{R}^{n \times n}$, such that $H_\chi(v)$ is positive definite for all $v \in V$. Under these conditions $H_\chi(v)$ defines an inner product in the tangent space in any point $\chi(v) \in M$ by equation (2.3).
The length of a differentiable curve γ on M from $\gamma(a)$ to $\gamma(b)$ is defined as $\int_{[a,b]} ||\frac{\partial \gamma}{\partial t}||_{\gamma(t)} dt$, where the norm $||\ ||_{\gamma(t)}$ comes from the inner product on the respective tangent space. Now, the Riemannian distance between two points θ and $\tilde{\theta} \in M$ is defined as the infimum over all lengths of differentiable curves on M with these two points as starting point and end point, respectively. If $\tilde{\theta}$ is not too far away from θ, then the infimum is in fact a minimum; the (unique) curve of shortest length, connecting the two points, is called a geodesic. One can consider the usual Euclidean geometry as a special case of Riemannian geometry, where the metric is just defined by $H_{id}(\theta) = I_n$ for all θ, and the geodesics are the straight lines.
In order to calculate geodesics, we need the notion of the Levi Civita connection $\nabla_X Y$, which is a rule for derivation of a vector field Y on the manifold with respect to the vector field X. This rule depends on the inner product that has been defined on the tangent spaces. A curve γ turns out to be a geodesic, if the derivation of its tangent field $\frac{\partial \gamma}{\partial t}$ with respect to itself is zero. This yields a set of differential equations, which has to be solved in order to find the geodesics. In general, it will be difficult to solve these equations. To simplify the equations, we need to find a parametrization $\chi : V \to M$, such that H_χ takes a simple form. For example, if H_χ has the form:

$$H_\chi(v_1, \ldots, v_n) = \begin{pmatrix} h(v) & 0 \\ 0 & c \end{pmatrix} \tag{2.5}$$

for some constant $c \in \mathbb{R}^+$, then the curves $v_n \mapsto \chi(v_1, \ldots, v_n)$ are in fact geodesics, because from the properties of the Levi Civita connection it follows that $\nabla_{\frac{\partial}{\partial v_n}} \frac{\partial}{\partial v_n} = 0$.
As we shall use frequently these properties of the Levi Civita-connection, we list them here. The Levi Civita-connection maps every pair of C^∞ vector fields X, Y on M to another C^∞ vector field on M, denoted by $\nabla_X Y$. This map is bilinear, and is in fact completely determined by the following[1]:
If $\chi(v_1, \ldots, v_n)$ is a parametrization of M, then

1. A C^∞-vectorfield X on M can be written as
   $X = f_1(v)\frac{\partial}{\partial v_1} + \ldots + f_n(v)\frac{\partial}{\partial v_n}$, where $f_i(v)$ are C^∞-differentiable functions

2. $\nabla_{f(v)\frac{\partial}{\partial v_i}} \frac{\partial}{\partial v_j} = f(v) \nabla_{\frac{\partial}{\partial v_i}} \frac{\partial}{\partial v_j}$

3. $\nabla_{\frac{\partial}{\partial v_i}} f(v)\frac{\partial}{\partial v_j} = \frac{\partial f(v)}{\partial v_i}\frac{\partial}{\partial v_j} + f(v)\nabla_{\frac{\partial}{\partial v_i}} \frac{\partial}{\partial v_j}$

[1] In differential geometrical methods in statistics, other connections than the Levi-Civita, are also important, see Amari (1987, 1990). These other connections are also derivation rules for a vector field Y with respect to a vector field X, with the same properties as the ones stated here for the Levi-Civita connection except for the fifth. However, we do not need these other connections.

4. $\nabla_{\frac{\partial}{\partial v_i}} \frac{\partial}{\partial v_j} = \nabla_{\frac{\partial}{\partial v_j}} \frac{\partial}{\partial v_i}$

5. $\frac{\partial}{\partial v_i}\langle \frac{\partial}{\partial v_j}, \frac{\partial}{\partial v_k}\rangle|_{\chi(v)} = \langle \nabla_{\frac{\partial}{\partial v_i}} \frac{\partial}{\partial v_j}, \frac{\partial}{\partial v_k}\rangle|_{\chi(v)} + \langle \frac{\partial}{\partial v_j}, \nabla_{\frac{\partial}{\partial v_i}} \frac{\partial}{\partial v_k}\rangle|_{\chi(v)}$.

From this, it follows that if $H_\chi = I_n$, or in general, if H_χ is constant on the manifold (the Euclidean case), then $\nabla_{\frac{\partial}{\partial v_i}} \frac{\partial}{\partial v_j} = 0$.
One can show that for any point $p \in M$, and for any tangent vector d in the tangent space M_p, there is exactly one geodesic $\gamma_{p,d}$, such that $\gamma_{p,d}(0) = p$, and $\frac{\partial \gamma_{p,d}}{\partial t}(0) = d$. If the geodesic $\gamma_{p,d}$ is defined on the interval $[0,t]$, then the geodesic $\gamma_{p,td}$ is defined on the interval $[0,1]$. One gets

$$\gamma_{p,td}(1) = \gamma_{p,d}(t).$$

Hence, there is a neighbourhood D of zero in the tangent space M_p, such that the geodesics $\gamma_{p,d}$ are all defined on $[0,1]$ for all $d \in D$. Because of this, one can construct a map from that neighbourhood D of zero in the tangent space M_p in the point p to the manifold M:

$$\exp_p : D \to M \quad ; \exp_p(d) = \gamma_{p,d}(1) \tag{2.6}$$

This map $\exp_p$ turns out to be C^∞, and is called the *normal map in the point p.*

2.2 Further aspects of Riemannian geometry

In this section, we have gathered topics of Riemannian differential geometry that are of importance for chapter three. Main reference is Helgason (1978).

2.2.1 Topological aspects

Let M be an n-dimensional Riemannian manifold[2]. M has a topology as a differential manifold[3], but M has also a topology as a metric space with distance function d, as explained in the first section. These topologies coincide (see e.g. corollary I.9.5, Helgason). M is connected (by definition), locally compact and separable (Proposition I.9.6, Helgason). The local situation can be more precisely described by a property that we shall refer to as the *normal property* :

- For every point $p \in M$ there exists a compact ball $B_{r_p}(p) = \{q \in M | d(p,q) \leq r_p\}$, such that in every smaller ball with center p any two points can be connected by exactly one curve in that smaller ball with a length equal to the distance of these two points (Helgason, Theorem I.9.9). (2.7)

Furthermore, any curve connecting two points with a length equal to the distance of these points, is a geodesic (Helgason, Lemma I.9.8).
This implies that from any point $q \in B_{r_p}(p)$ with distance $d(p,q) < \eta$ to p any other point $\tilde{q}$ in the ball $B_{r_p - \eta}(q)$ can be reached by a unique geodesic having a length $d(q,\tilde{q})$.

[2]For a precise definition, the reader is referred to the books, mentioned above.
[3]For instance, if $M = AR_n$, M has a topology as subset of $\mathbf{R}^n$.

Hence, the normal mapping $\exp_q$ is defined on a ball with radius $r_p - \eta$ around 0 in the tangent space M_q. There it is an invertible map to $B_{r_p-\eta}(q) \subset M$. Because of this, $B_{r_p}(p)$ is called a *normal neighbourhood* of each of its points.
If the normal map $\exp_p$ is defined on the whole tangent space for all $p \in M$, i.e., if all the geodesics with maximal domain are defined on the whole $\mathbb{R}$, the Riemannian manifold is called *geodesically complete*. According to Helgason, Theorem I.10.3, M is geodesically complete if and only if it is metrically complete, i.e. if all the Cauchy sequences in M converge in M. In this theorem, it is also stated that M is complete if and only if any bounded closed subset of M is compact.

2.2.2 Charts

On an n-dimensional manifold M we can *locally* define coordinates, i.e. every point $q \in M$ has a neighbourhood that is homeomorphic with an open subset of $\mathbb{R}^n$. More precisely, M can (by definition) be covered by open sets $\tilde{U}_\alpha$, such that there exist homeomorphisms $x_\alpha : \tilde{U}_\alpha \to V_\alpha$ from these sets to open subsets V_α of $\mathbb{R}^n$, and such that these homeomorphisms are mutually compatible. By "mutually compatible" we mean that if $x_\beta : \tilde{U}_\beta \to V_\beta$ and $x_\alpha : \tilde{U}_\alpha \to V_\alpha$ are such n-tuples of coordinates, then the transformation from one to the other n-tuple on $\tilde{U}_\alpha \cap \tilde{U}_\beta$ should be C^∞-differentiable:

$$x_\beta \circ x_\alpha^{-1} : x_\alpha(\tilde{U}_\alpha \cap \tilde{U}_\beta) \to V_\beta \text{ is } C^\infty;$$

$$x_\alpha \circ x_\beta^{-1} : x_\beta(\tilde{U}_\alpha \cap \tilde{U}_\beta) \to V_\alpha \text{ is } C^\infty.$$

The couple $(\tilde{U}_\alpha, x_\alpha : \tilde{U}_\alpha \to V_\alpha)$ is called a *chart*. The sets $\tilde{U}_\alpha$ will be called *coordinate neighbourhoods*. Instead of expressing everything in the "coordinates" $x_\alpha : \tilde{U}_\alpha \to V_\alpha$, we can as well work with their inverses, the homeomorfisms $\psi_\alpha = x_\alpha^{-1} : V_\alpha \to \tilde{U}_\alpha$. We shall call these inverses "parametrizations". We did so in the introduction preceding this section. We have: $\psi_\alpha(\hat{w}_1, \cdots, \hat{w}_n) = p$ if and only if $x_\alpha(p) = (\hat{w}_1, \cdots, \hat{w}_n)$. If M is an open subset of $\mathbb{R}^n$, then the partial derivative $\frac{\partial \psi_\alpha}{\partial w_i}_{|\hat{w}}$ corresponds with the tangent vector $\frac{\partial}{\partial x_{\alpha,i}}$ in the tangent space M_p.
Because M is separable, it is possible to cover M by a countable number of open sets U_α, which have a compact closure in open coordinate neighbourhoods $\tilde{U}_\alpha$. On these open sets U_α we have in fact two natural metrics: the Riemannian metric $\boldsymbol{d}$, but also the metric $\tilde{\boldsymbol{d}}$, induced by the Euclidean norm on V_α:

$$\tilde{\boldsymbol{d}}(p,q) = ||x_\alpha(p) - x_\alpha(q)||_{\text{Eucl}}.$$

The following lemma 2.1 discloses something about the relation between the two metrics on U_α.

Lemma 2.1 Let $(\tilde{U}_\alpha, x_\alpha : \tilde{U}_\alpha \to V_\alpha \subset \mathbb{R}^n)$ be a chart of the Riemannian manifold M. Let K be a compact subset of $\tilde{U}_\alpha$. Then, there exist positive numbers ε_K and M_K, such that

$$\boldsymbol{d}(p,q) < M_K ||x_\alpha(p) - x_\alpha(q)||_{\text{Eucl}} \quad \text{if} \quad ||x_\alpha(p) - x_\alpha(q)||_{\text{Eucl}} < \varepsilon_K \tag{2.8}$$

$$||x_\alpha(p) - x_\alpha(q)||_{\text{Eucl}} < M_K \boldsymbol{d}(p,q) \quad \text{if} \quad \boldsymbol{d}(p,q)) < \varepsilon_K \tag{2.9}$$

for all $p, q \in K$.

Proof: Because of the the normal property (2.7), mentioned above, we can find a positive number r_p for every point $p \in K$, such that

- the ball $B_{r_p}(p)$ is compact,
- every two points in it can be joined by exactly one curve inside the ball with length equal to the distance of these two points,
- and $B_{r_p}(p) \subset \tilde{U}_\alpha$.

Only a finite number of balls $B_{\frac{1}{2}r_{p_i}}(p_i)$ $(i = 1, \cdots, k)$ is needed to cover K. Then, also

$$K^1 = \bigcup_{i=1,\cdots,k} B_{r_{p_i}}(p_i)$$

is a compact subset of $\tilde{U}_\alpha$. The quadratic form defining the Riemannian norms on the tangent spaces M_q of $\tilde{U}_\alpha$, is a C^∞ function of its coordinates in V_α. Consequently, on K^1 these norms $||\ ||_q$ on the tangent spaces are uniformly equivalent with the Euclidean norm $||\ ||_{\text{Eucl}}$, i.e. there exist positive numbers C and D, such that

$$C||\sum_{i=1}^{n} v_i \frac{\partial}{\partial x_{\alpha,i}}||_q \leq ||v||_{\text{Eucl}} \leq D||\sum_{i=1}^{n} v_i \frac{\partial}{\partial x_{\alpha,i}}||_q \text{ for all } q \in K^1 \text{ and } v \in \mathbb{R}^n.$$

Let $\varepsilon_K = \frac{1}{2}\min_{i=1,\cdots,k} r_{p_i}$. For any $p \in K$, for which $\boldsymbol{d}(p,q) < \varepsilon_K$ holds, we have that both points p and q are in the same $B_{r_{p_j}}(p_j)$. Let $\gamma : [0,1] \to M; \gamma(0) = p; \gamma(1) = q$ be the unique curve in $B_{r_{p_j}}(p_j)$ with length equal to $\boldsymbol{d}(p,q)$. Then we have:

$$x_\alpha \circ \gamma(t) = (\gamma_1(t), \cdots, \gamma_n(t))\ ;\ \gamma'(t) = \sum_{i=1}^{n} \gamma_i'(t) \frac{\partial}{\partial x_{\alpha,i}} \text{ and}$$

$$||x_\alpha(p) - x_\alpha(q)||_{\text{Eucl}} \leq \int_0^1 ||(x_\alpha \circ \gamma)'(t)||_{\text{Eucl}} dt \leq \int_0^1 D||\gamma'(t)||_q dt = D\boldsymbol{d}(p,q).$$

The proof that we can find the positive numbers ε_K and M_K, such that (2.8) is satisfied, proceeds in the same way, but is simpler. □

2.2.3 Derivations. Derivative of the normal mapping

We give some more abstract definitions that we need in some proofs in chapter three. Let M and N be respectively n- and k-dimensional differentiable manifolds.

- A function $f : M \to \mathbb{R}$ is called C^∞ if for all charts $(\tilde{U}_\alpha, x_\alpha : \tilde{U}_\alpha \to V_\alpha \subset \mathbb{R}^n)$ the function $f \circ x_\alpha^{-1} : V_\alpha \to \mathbb{R}$ is C^∞. The set of the C^∞-functions $f : M \to \mathbb{R}$ is a linear space, even an algebra, denoted by $\mathcal{F}$ (or $\mathcal{F}_M$).

- The tangent vectors in the tangent space M_p can be identified with the *derivations at* $p \in M$. A derivation d at p is a linear map $d : \mathcal{F}_M \to \mathbb{R}$, such that

$$d(fg) = f(p)d(g) + g(p)d(f) \text{ for all } f, g \in \mathcal{F}_M \text{ (product rule).}$$

For instance, if $d = \frac{\partial}{\partial x_{\alpha,i}|_p}$ then d defines the derivation

$$d(f) = \frac{\partial f \circ x_\alpha^{-1}}{\partial v_i}(x_\alpha(p)) \text{ for all } f \in \mathcal{F}_M.$$

- A map $\psi : N \to M$ is called C^∞ if for all C^∞-functions $f : M \to \mathbb{R}$ also the composition $f \circ \psi : N \to \mathbb{R}$ is C^∞.

- Such a C^∞-map $\psi : N \to M$ induces a linear map $\psi_{*|p} : N_p \to M_{\psi(p)}$ between the tangent spaces in the original point and the image point, defined as follows: if d is a tangent vector in N_p (or a derivation at p), we define the derivation $\psi_{*|p}$ at $\psi(p)$ by:

$$\psi_{*|p}(d) : \mathcal{F}_M \to \mathbb{R};\ \psi_{*|p}(d)(f) = d(f \circ \psi).$$

This linear map is called the *derivative* of ψ at p.

- A C^∞-vector field X on M can be defined as a *derivation of the algebra* $\mathcal{F}_M$, i.e. a linear map $X : \mathcal{F}_M \to \mathcal{F}_M$, such that

$$X(fg) = fXg + gXf.$$

One can see the vector field X just as a bunch of tangent vectors $X(p)$ in the points p of the manifold M, such that for every C^∞ function f on M also its derivative Xf in the directions $X(p)$ is again a C^∞ function on M. On a chart $(\tilde{U}_\alpha, x_\alpha : \tilde{U}_\alpha \to V_\alpha)$ a C^∞ vector field X can be written

$$X(p) = f_1(p)\frac{\partial}{\partial x_{\alpha,1}|_p} + \cdots + f_n(p)\frac{\partial}{\partial x_{\alpha,n}|_p} \text{ for all } p \in \tilde{U}_\alpha$$

for some C^∞-functions $f_1, \cdots, f_n \in \mathcal{F}_M$.
The set of all C^∞-vector fields on M will be denoted by $\mathcal{D}M$.

- The *Lie derivative* of the C^∞ vector field X with respect to the C^∞ vector field Y is given by:

$$[X, Y]f = X(Yf) - Y(Xf) \text{ for all } f \in \mathcal{F}_M.$$

For coordinate vector fields $\frac{\partial}{\partial x_{\alpha,1}}, \cdots, \frac{\partial}{\partial x_{\alpha,n}}$ one has:

$$[\frac{\partial}{\partial x_{\alpha,i}}, \frac{\partial}{\partial x_{\alpha,j}}] = 0 \text{ for all } i, j.$$

- The Levi-Civita-connection ∇ is a bilinear map from $\mathcal{D}M \times \mathcal{D}M$ into $\mathcal{D}M$ satisfying the properties 2, 3, 4 and 5, as we have stated in the first section. The fourth property can be rewritten as:

$$4.\ [X,Y] = \nabla_X Y - \nabla_Y X \text{ for all } X, Y \in \mathcal{D}M.$$

In the first section we have introduced the normal mapping in a point $q \in M$. It is a map from an open subset D around 0 in the tangent space M_q to a neighbourhood $\tilde{U}$ around the point q in M:

$$\exp_q : D \to \tilde{U}.$$

We denote its *derivative* at the point $d \in M_q$ as $\exp_{q*|d}$. This derivative is a mapping from $(M_q)_d$ to $M_{\exp_q(d)}$. We usually identify the spaces $(M_q)_d$ and M_q. It can be (easily) proved that the derivative of the normal mapping in q at $0 \in M_q$ is the identity:

$$\exp_{q*|0}(w) = w \qquad \forall w \in M_q = (M_q)_0.$$

2.2.4 Special types of charts

A very special type of charts will be of interest: *normal charts.* Their construction is as follows.
Let p be a point in the Riemannian manifold M, and let $e_1, \cdots, e_n$ be a basis in the tangent space M_p. There exists a ball $B_{r_p}(p)$ around the point p, such that the normal property (2.7) holds. Let $U_{p\mathrm{n}}$ be an open subset around p in $B_{r_p}(p)$. Then, $U_{p\mathrm{n}}$ is a normal neighbourhood and there exists an open set $V_{p\mathrm{n}}$ around 0 in $\mathbb{R}^n$, such that the function $\psi_{p\mathrm{n}}$,

$$\psi_{p\mathrm{n}} : V_{p\mathrm{n}} \to U_{p\mathrm{n}},$$

$$\psi_{p\mathrm{n}}(x_1, \cdots, x_n) = \exp_p(\sum_{i=1}^{n} x_i e_i)$$

is a diffeomorfism. Any $q \in U_{p\mathrm{n}}$ can be written as $q = \psi_{p\mathrm{n}}(x_1, \cdots, x_n)$, where $(x_1, \cdots, x_n)' \in V_{p\mathrm{n}}$. Then, we define the *normal chart* $(U_{p,n}, x_{p\mathrm{n}e,w})$ *around* p, *using tangent base* $e_1, \cdots, e_n$ *and vector* $w \in \mathbb{R}^n$ as follows:

$$\begin{aligned} x_{p\mathrm{n}e,w} : U_{p\mathrm{n}} &\to V_{p\mathrm{n}} \\ x_{p\mathrm{n}e,w}(q) &= (w_1, \cdots, w_n) + (x_1, \cdots, x_n) \text{ if } q = \exp_p(\sum_{i=1}^{n} x_i e_i). \end{aligned} \tag{2.10}$$

On this normal chart any geodesic γ with $\gamma(0) = p$ and $\dot{\gamma}(0) = \sum_{i=1}^{n} x_i e_i \in M_p$ has coordinates

$$x_{p\mathrm{n}e,w}(\gamma(t)) = x_{p\mathrm{n}e,w}(\exp_p(t \sum_{i=1}^{n} x_i e_i) = \vec{w} + t\vec{x}. \tag{2.11}$$

Hence, in these coordinates, geodesics through the point p are just (pieces of) straight lines.

We give a well known fact that will be used as a tool to construct another type of charts: charts for which the corresponding parametrization is a *pencil of geodesics* (This construction can be found in chapter three (example 3.10 and continuation)).
Let M be an n-dimensional Riemannian manifold. Let I be an open interval in $\mathbb{R}$, and J be an open subset of $\mathbb{R}^k$ $(1 \leq k \leq n-1)$. A C^∞-map $f : I \times J \to M$ is called a *pencil of geodesics*if for all $s \in J$ the curve

$$\gamma : I \to M; \gamma(t) = f(t,s) \tag{2.12}$$

is a geodesic parametrized, according to arc length. Let us study the pencil of geodesics f. We define the C^∞-vector fields F and Z_i, $1 \leq i \leq k$ on the image of f:

$$F := f_*(\frac{\partial}{\partial t}) = \sum_{i=1}^{n} \frac{\partial x_{\alpha\, i} \circ f}{\partial t} \frac{\partial}{\partial x_{\alpha\, i}};$$

$$Z_i := f_*(\frac{\partial}{\partial s_i}) = \sum_{i=1}^{n} \frac{\partial x_{\alpha\, i} \circ f}{\partial s_i} \frac{\partial}{\partial x_{\alpha\, i}}$$

on any chart $(\tilde{U}_\alpha, x_\alpha)$. For the geodesic γ, given in (2.12), we have: $\dot{\gamma}(t) = F$, hence $\nabla_F F = 0$. Since the curve γ is also parametrized, according to arc length, we have: $\langle F, F\rangle_{f(t,s)} = 1$ for all $(t,s) \in I \times J$. Because

$$\frac{\partial^2 x_{\alpha\, i} \circ f}{\partial s_i \partial t} = \frac{\partial^2 x_{\alpha\, i} \circ f}{\partial t \partial s_i} \text{ for all } i,\ 1 \leq i \leq n$$

on any chart $(\tilde{U}_\alpha, x_\alpha)$, we get:

$$Z_i F = F Z_i, \text{ or } [Z_i, F] = 0 \text{ so } \nabla_{Z_i} F = \nabla_F Z_i.$$

Well known fact: The inner product of the vector fields Z_i with the vector field F is constant along all geodesics γ, defined in (2.12), i.e. the function:

$$t \mapsto \langle Z_i, F\rangle_{f(t,s)} \quad (t \in I, s \in J) \text{ is constant on } I \text{ for all } s \in J.$$

Proof:

$$\frac{\partial}{\partial t}\langle Z_i, F\rangle_{f(t,s)} = F\langle Z_i, F\rangle_{f(t,s)} = \langle \nabla_F Z_i, F\rangle_{f(t,s)} + \langle Z_i, \nabla_F F\rangle_{f(t,s)} =$$

$$= \langle \nabla_{Z_i} F, F\rangle_{f(t,s)} + 0 = \frac{1}{2} Z_i \langle F, F\rangle_{f(t,s)} = \frac{\partial}{\partial s_i}(1) = 0.\square$$

2.2.5 Tensor fields and distributions

Let $\mathcal{F}_M$ be the linear space of all C^∞ functions from M to $\mathbb{R}$. Let $\mathcal{D}M$ be the set of all C^∞- vector fields on M. Then, $\mathcal{D}M$ is a linear vector space, but it is also an $\mathcal{F}_M$-module. For every C^∞ function $f \in \mathcal{F}_M$ and every C^∞-vector field $X \in \mathcal{D}M$, we also have that their product fX is a C^∞ vector field: $fX \in \mathcal{D}M$.
A (smooth)*distribution* is a submodule of $\mathcal{D}M$, generated by a finite or (possibly) infinite

number of C^∞ vector fields. If the distribution $\mathcal{D}$ is generated by $X_1, \cdots, X_k$, then in any tangent space M_p we have the (at most k- dimensional) subspace $\mathcal{D}(p) = \boldsymbol{span}\{X_1(p), \cdots, X_k(p)\}$. We say that the distribution $\mathcal{D}$ is constant dimensional, if there is a number m, such that $\dim \mathcal{D}(p) = m$ for all $p \in M$. $\mathcal{D}M$ itself is a constant dimensional distribution.
Also $\mathcal{D}M^k = \mathcal{D}M \times \cdots \times \mathcal{D}M$ (k times) is an $\mathcal{F}_M$-module. We denote $\mathcal{F}_M$ some times as $\mathcal{D}M^0$.
A *tensor field of type* (r, s) is a map $L : \mathcal{D}M^r \to \mathcal{D}M^s$, which is r-linear over $\mathcal{F}_M$, i.e.

$$L(X_1, \cdots, X_i + Y_i, \cdots, X_r) = L(X_1, \cdots, X_i, \cdots, X_r) + L(X_1, \cdots, Y_i, \cdots, X_r);$$

$$L(X_1, \cdots, fX_i, \cdots, X_r) = fL(X_1, \cdots, X_i, \cdots, X_r) \quad (f \in \mathcal{F}_M) \quad (1 \leq i \leq r).$$

A well known tensor field of type $(2,0)$ is the Riemannian inner product. The curvature tensor

$$R(X,Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z \tag{2.13}$$

is a tensor field of type $(3,1)$. The Riemannian connection $(X,Y) \mapsto \nabla_X Y$ is *not* a tensor field of type $(2,1)$, because $\nabla_X fY = (Xf)Y + f\nabla_X Y$ for all $f \in \mathcal{F}_M$. Any function $f \in \mathcal{F}_M$ induces a tensor field $\mathrm{d}f$ of type $(1,0)$ by:

$$\mathrm{d}f(X) = Xf \quad (X \in \mathcal{D}M).$$

A tensor field of type $(1,0)$ is called a *differential form.*
A very important property of a tensor field of type (r,s) is that its value in the point $p \in M$ is completely determined by the values of the vector fields in the point p; it is not necessary to know the vector fields in a neighbourhood of p. We denote this as follows:

$$L(X_1, \cdots, X_r)(p) = L(X_1(p), \cdots, X_r(p)).$$

A tensor field L of type $(2,s)$ is called *symmetric* if $L(X,Y) = L(Y,X)$. The Riemannian inner product is a symmetric tensor field. Any other symmetric tensor field V of type $(2,0)$ gives rise to a tensor field $\tilde{V}$ of type $(1,1)$ by using the duality property of the Riemannian inner product:

$$V(X,Y)(p) = \langle \tilde{V}(X), Y \rangle_p.$$

In fact, $\tilde{V}$ is a symmetric linear map from every tangent space M_p into itself. With some abuse of notation, we shall often denote the symmetric tensor field V of type $(2,0)$, and the tensor field $\tilde{V}$ of type $(1,1)$ by the same symbol V.

2.2.6 Involutive distributions, Submanifolds and Curvature

A distribution $\mathcal{D}$ is called **involutive** if for all vector fields $X, Y \in \mathcal{D}$ also the vector field $[X,Y] \in \mathcal{D}$ holds. If the distribution $\mathcal{D}$ on the manifold M is constant dimensional and involutive, then the theorem of Frobenius (see e.g. Kobayashi and Nomizu (1963) or Nijmeijer and Van der Schaft, Corollary 2.43) states that we can cover M by charts $(\tilde{U}_\alpha, x_\alpha : \tilde{U}_\alpha \to V_\alpha)$, such that

$$\mathcal{D}(p) = \boldsymbol{span}\{\frac{\partial}{\partial x_{\alpha 1}}, \cdots, \frac{\partial}{\partial x_{\alpha\, m}}\} \text{ for all } p \in \tilde{U}_\alpha.$$

In that case, for any $p \in \tilde{U}_\alpha$ there is a neighbourhood U, such that the subset N of M,

$$N = \{q \in U |\ x_{\alpha\, m+i}(q) = x_{\alpha\, m+i}(p)\ \ \forall i, 1 \leq i \leq n-m\}, \tag{2.14}$$

is a submanifold of M, and the tangent space N_p is equal to $\mathcal{D}(p)$. We shall say that such a set of charts is *adapted to the distribution* $\mathcal{D}$.
A submanifold N of M that satisfies $N_p = \mathcal{D}(p)$ for all $p \in N$, is called an *integral manifold* of the distribution $\mathcal{D}$. Every integral manifold of $\mathcal{D}$ is of the form (2.14).
A **submanifold** N of an n-dimensional Riemannian manifold M has its own Riemannian structure. We get the Levi-Civita-connection $\tilde{\nabla}$ of N by projecting the vector $\nabla_X Y(p)$ onto the tangent space N_p for all vector fields X, Y on N. The relation between the two connections is:

$$\nabla_X Y(p) = \tilde{\nabla}_X Y(p) + V(X,Y)(p) \qquad (p \in N)$$

where V is a map from $\mathcal{D}N \times \mathcal{D}N \to \mathcal{D}N^\perp \subset \mathcal{D}M$[4]. The map V is $\mathcal{F}_N$-bilinear, symmetric, and is called the *second fundamental form of N in M.*

- If the dimension of N is $n-1$, i.e. if N is a hypersurface of M, then for any point $p \in N$ there is a neighbourhood U in N of p, and a vector field ξ on U, which is normal to N everywhere. Furthermore, it has e.g. length one everywhere, and it is perpendicular to N_q for all $q \in U$. Moreover, for all vector fields X, Y on U we have:

$$\langle Y, \xi\rangle = 0, \text{ hence } 0 = X\langle Y, \xi\rangle = \langle \nabla_X Y, \xi\rangle + \langle Y, \nabla_X \xi\rangle, \text{ hence}$$
$$V(X,Y) = \langle \nabla_X Y, \xi\rangle \xi = -\langle Y, \nabla_X \xi\rangle \xi.$$

  Thus, the second fundamental form is completely determined by $X \mapsto \nabla_X \xi$. As $1 = \langle \xi, \xi\rangle$, we have $0 = X\langle \xi, \xi\rangle = 2\langle \nabla_X \xi, \xi\rangle$, consequently, the map $X(q) \mapsto \nabla_X \xi(q)$ is a map of N_q into N_q for all $q \in U$. This map is called *Weingarten's* map. Because of the symmetry of V, we have $\langle Y, \nabla_X \xi\rangle = \langle X, \nabla_Y \xi\rangle$, hence Weingarten's map is symmetric.

- A submanifold N of M is called *totally geodesic* if all the geodesics of N as Riemannian manifold are also geodesics of M. A submanifold N of M is totally geodesic if and only if the Riemannian connection of N is equal to the Riemannian connection of M, i.e. the second fundamental form is zero. This is easily seen as follows. If the second fundamental form is zero, then it follows from $\tilde{\nabla}_X X = 0$ that $\nabla_X X = 0$, and N is totally geodesic. If N is totally geodesic, then we have for all *geodesic* tangent fields X on N that $\tilde{\nabla}_X X = \nabla_X X = 0$, consequently $V(X,X) = 0$. Because V is $\mathcal{F}_N$-bilinear, we have for all vector fields X on N that $V(X,X) = 0$. Because V is symmetric, it also implies that $V(X,Y) = 0$ for all vector fields X, Y on N, thus the second fundamental form of N is zero.
  A constant dimensional distribution $\mathcal{D}$ is called *totally geodesic* if for all $X, Y \in \mathcal{D}$ also $\nabla_X Y \in \mathcal{D}$ holds. From what has been remarked about involutive distributions and totally geodesic submanifolds, it is clear that a constant dimensional distribution is totally geodesic if and only if it is involutive and all its integral manifolds are totally geodesic.

[4]See e.g. Kobayashi and Nomizu, Proposition VII.3.1.

Curvature. The curvature tensor of type $(3,1)$, denoted by $R(X,Y)Z$ for $X,Y,Z \in \mathcal{D}M$, was already introduced in equation (2.13). In the same way as there is a relation between symmetric tensor fields of type $(2,0)$, and tensor fields of type $(1,1)$, there is a relation between tensor fields of type $(3,1)$, and certain tensor fields of type $(4,0)$. The tensor field K of type $(4,0)$ is defined by:

$$K(U,Z,X,Y)(p) = \langle R(X,Y)Z,U\rangle_p, \tag{2.15}$$

and is also called curvature tensor. It has the following (anti)symmetric properties:

1.
$$K(U,Z,X,Y) = -K(U,Z,Y,X);$$

2.
$$K(U,Z,X,Y) = K(X,Y,U,Z);$$

3. if both pairs $X(p),Y(p)$ and $U(p),Z(p)$ are orthonormal and both pairs span the same subspace of M_p

$$\text{then } K(U,Z,U,Z) = K(X,Y,X,Y).$$

The third property enables us to define the *sectional curvature* $K(V)$ *of the two-dimensional subspace* V *of the tangent space* M_p by:

$$\text{if } \textit{span}\{X(p),Y(p)\} = V, \text{ then } K(V) = \frac{K(X,Y,X,Y)(p)}{\langle X,X\rangle_p\langle Y,Y\rangle_p - \langle X,Y\rangle_p^2}. \tag{2.16}$$

Using the third property, it is easy to show that this definition does not depend on the chosen base $X(p),Y(p)$ of V. In the book of Helgason the rigorous connection with the intuitive interpretation of curvature is described: if $B_{r_p}(p)$ is a ball with the normal property (2.7) and D_r is a ball (disk) with radius $r < r_p$ with center 0 in V, then one can compare the area of this disk with the area of its image under the normal map: $\exp_p : M_p \to M$. A more intuitive definition of curvature would be:

$$K(V) = \lim_{r\to 0} 12\frac{\text{area}(D_r) - \text{area}(\exp_p(D_r))}{r^2\text{area}(D_r)} \tag{2.17}$$

In Theorem I.12.2, Helgason states that both definitions of sectional curvature coincide. As a submanifold N has its own Riemannian structure, it also has its own curvature tensor. Gauss has compared the sectional curvature $\tilde{K}(V)$ of a two-dimensional subspace V of the tangent space $N_p \subset M_p$ in this Riemannian structure with the sectional curvature $K(V)$ in the Riemannian structure of M. In case N is a hypersurface with normal vector field ξ, then the formula Gauss derived can be expressed in terms of Weingartens' map $L : N_q \to N_q$; $L(X(q)) = \nabla_X\xi(q)$. Then, the relation is:

$$K(V) = \tilde{K}(V) - (\langle Lv,v\rangle\langle Lw,w\rangle - \langle Lv,w\rangle^2) \tag{2.18}$$

where v,w are a pair of orthonormal vectors spanning V [5].

[5] Kobayashi and Nomizu, Prop.VII.4.5.

2.3 Stochastic variables with values in a tangent bundle

As stated in the introduction, we want to generalize the usual concept of a random walk to that of a stochastic process on a Riemannian manifold. Hence, we need to study stochastic variables with values in a Riemannian manifold. In general, an n-dimensional Riemannian manifold M does not have an additive structure. Therefore, it is difficult to see what "expectation" and "variance" of such a variable would mean. The tangent spaces of M have an additive structure, because they are isomorphic with $\mathbb{R}^n$. We can introduce some kind of concept of conditional expectation and conditional variance for stochastic variables with values in the *tangent bundle* TM of M. The tangent bundle is a well known differential geometrical structure which can be defined as follows:

Let $((\tilde{U}_\alpha, x_\alpha : \tilde{U}_\alpha \to V_\alpha \subset \mathbb{R}^n))_{\alpha \in \Lambda}$ be a collection of charts that together covers the manifold M.
We define the *tangent bundle* TM of the manifold M by:

$$TM = \{(q,d) |\ \ q \in M\ ;\ d \in M_q\}. \tag{2.19}$$

The tangent bundle TM can also be made a topological space, and a differential manifold by defining:

- the projection $\pi : TM \to M\ ;\ \pi(q,d) = q$;
- the coordinate neighbourhoods $\pi^{-1}(V_\alpha)$;
- the "coordinates" $\tilde{x}_\alpha : \pi^{-1}(\tilde{U}_\alpha) \to \tilde{x}_\alpha(\pi^{-1}(\tilde{U}_\alpha)) \subset \mathbb{R}^{2n}$;

$$\tilde{x}_\alpha(q, \sum_{i=1}^{n} d_i \frac{\partial}{\partial x_{\alpha,i}}) = (x_\alpha(q), d), \text{ considered as homeomorphisms.}$$

In this way, the projection $\pi : TM \to M$ is continuous.

Let $(\Omega, \mathcal{A}, P)$ be a probability (and hence a measurable) space. Also the manifold M can be considered a measurable space, with its Borel sets as σ-algebra. As is easily verified, a mapping between these measurable spaces, $q : \Omega \to M$, is $\mathcal{A}$-measurable if and only if $x_\alpha(q) 1_{\{q \in \tilde{U}_\alpha\}}$ is $\mathcal{A}$-measurable for all charts $(\tilde{U}_\alpha, x_\alpha)$[6].
Like M, also its tangent bundle TM is a measurable space with the Borel sets as σ-algebra. A mapping $(q,d) : \Omega \to TM$ is $\mathcal{A}$-measurable if and only if $\tilde{x}_\alpha(q,d) 1_{\{(q,d) \in \pi^{-1}(\tilde{U}_\alpha\}}$ is $\mathcal{A}$-measurable. If $(q,d) : \Omega \to TM$ is $\mathcal{A}$-measurable, then its projection q on M is $\mathcal{A}$-measurable, and also $||d||_q$ is an $\mathcal{A}$-measurable function.

[6]Here we use that M is covered by a countable number of sets $\tilde{U}_\alpha$.

Conditional Expectation. Assumptions: Let $(q,d):\Omega\to TM$ be $\mathcal{A}$-measurable. Let its projection $q:\Omega\to M$ be $\mathcal{B}$-measurable, where $\mathcal{B}$ is a sub-σ-algebra of $\mathcal{A}$. Suppose furthermore that

$$\boldsymbol{E}(||d||_q 1_{\{q\in K\}})<\infty \text{ for all compact } K\subset M.$$

We are going to define the $\mathcal{B}$-measurable mapping

$$(q,\boldsymbol{E}^{\mathcal{B}}d):\Omega\to TM.$$

Let $(\,(\tilde{U}_\alpha,x_\alpha:\tilde{U}_\alpha\to V_\alpha\subset\mathbb{R}^n)\,)_{\alpha\in\Lambda}$ be a collection of charts, and let $(\tilde{U}_\alpha)_{\alpha\in\Lambda}$ be a collection of open sets, such that U_α has compact closure in the coordinate neighbourhood $\tilde{U}_\alpha$ for all $\alpha\in\Lambda$, and such that the U_α cover M. If $q\in U_\alpha\cap U_\beta$, then we have two bases in M_q: $(\frac{\partial}{\partial x_{\alpha,i}})_{1\le i\le n}$, and $(\frac{\partial}{\partial x_{\beta,j}})_{1\le j\le n}$. Their relation is given by:

$$\frac{\partial}{\partial x_{\beta,j}}=\sum_{i=1}^{n}(Dx_\alpha\circ x_\beta^{-1})_{ij}\frac{\partial}{\partial x_{\alpha,i}}.$$

Consequently, if $d\in M_q$, we can write:

$$d=\sum_{i=1}^{n}\delta_{\alpha,i}\frac{\partial}{\partial x_{\alpha,i}}=\sum_{j=1}^{n}\delta_{\beta,j}\frac{\partial}{\partial x_{\beta,j}},$$

where

$$\delta_{\alpha,i}=\sum_{j=1}^{n}(Dx_\alpha\circ x_\beta^{-1})_{ij}\delta_{\beta,j}.$$

Note that $\{q\in U_\alpha\}\in\mathcal{B}$. The second coordinate d of the map $(q,d):\Omega\to TM$ satisfies:

$$d1_{\{q\in U_\alpha\}}=\sum_{i=1}^{n}\delta\alpha,i(q(\omega),\omega)1_{\{q\in U_\alpha\}}\frac{\partial}{\partial x_{\alpha,i}}.$$

Using the same reasoning as in lemma 2.1 (uniform equivalence of Riemannian norms with the usual Euclidean norm on the closure of U_α), we see that the functions $\delta_{\alpha,i}(q(\omega),\omega)1_{\{q\in U_\alpha\}}$ are not only $\mathcal{A}$-measurable, but also integrable.
Now we **define**:

$$\boldsymbol{E}^{\mathcal{B}}d1_{\{q\in U_\alpha\}}:=\sum_{i=1}^{n}\boldsymbol{E}^{\mathcal{B}}(\delta_{\alpha,i}(q(\omega),\omega)1_{\{q\in U_\alpha\}})\frac{\partial}{\partial x_{\alpha,i}}.$$

This is a good definition, because $Dx_\alpha\circ x_\beta^{-1}(x_\beta(q))$ is continuous in q, hence a $\mathcal{B}$-measurable function. □

Conditional Variance. Let the same assumptions hold as in the preceding definition. We **define** $(q,\boldsymbol{Var}^{\mathcal{B}}(d))$ as the following stochastic symmetric bilinear map from $\mathcal{D}M\times\mathcal{D}M$ to $\mathbb{R}$:

$$\boldsymbol{Var}^{\mathcal{B}}(d)(X,Y)=\boldsymbol{E}^{\mathcal{B}}(\langle d-\boldsymbol{E}^{\mathcal{B}}d,X(q)\rangle_q\langle d-\boldsymbol{E}^{\mathcal{B}}d,Y(q)\rangle_q),\tag{2.20}$$

where X, Y are C^∞-vector fields on M.
A simple verification shows that this is a good definition. The usual rule for variances:

$$\boldsymbol{Var}^{\mathcal{B}}(d)(X,Y) = \boldsymbol{E}^{\mathcal{B}}(\langle d, X(q)\rangle_q \langle d, Y(q)\rangle_q) - (\langle \boldsymbol{E}^{\mathcal{B}} d, X(q)\rangle_q \langle \boldsymbol{E}^{\mathcal{B}} d, Y(q)\rangle_q)$$

is valid. For any $X, Y \in \mathcal{D}M$, $\boldsymbol{Var}^{\mathcal{B}}(d)(X,Y) : \Omega \to \mathbf{R}$ is a $\mathcal{B}$-measurable map. In the same way as for a symmetric tensor field of type $(2,0)$, we can consider $\boldsymbol{Var}^{\mathcal{B}}(d)$ as a semi-positive definite symmetric linear map from M_q into M_q. □

We end this section with a small lemma that will be needed in one of the propositions of this chapter:

Lemma 2.2 Let $(\Omega, \mathcal{A}, P)$ be a measurable space, and M a Riemannian manifold. Let $(q_t)_{t\in\mathbf{N}}$ be a sequence of $\mathcal{A}$-measurable mappings $q_t : \Omega \to M$. Then, the set $\{\lim_{t\uparrow\infty} q_t$ exists in $M\}$ is $\mathcal{A}$-measurable. Let $h : \Omega \to M$ be another $\mathcal{A}$-measurable map. Then, also the map $q : \Omega \to M$, defined by:

$$q = \lim_{t\uparrow\infty} q_t, \text{ if this limit exists in M; if not } q = h$$

is $\mathcal{A}$-measurable.

Proof: Take a countable set of coordinate neighbourhoods $\tilde{U}_\alpha$ (together with coordinates $x_\alpha : \tilde{U}_\alpha \to V_\alpha \subset \mathbf{R}^n$) which cover M. Then,

$$\{\lim_{t\uparrow\infty} q_t \text{ exists in} M\} = \bigcup_\alpha \{\liminf x_\alpha(q_t 1_{\{q_t\in\tilde{U}_\alpha\}}) =$$

$$= \limsup x_\alpha(q_t 1_{\{q_t\in\tilde{U}_\alpha\}}) \wedge \limsup x_\alpha(q_t 1_{\{q_t\in\tilde{U}_\alpha\}}) \in V_\alpha\}.$$

This set is without any doubt measurable. It is now easy to prove the other statement as well. □

2.4 The asymptotic Fisher information of an AR(n) process

In statistics, the concept of Fisher information plays an important role. If the probability distribution p_θ of a stochastic variable X depends on a parameter θ in an open subset of $\mathbf{R}^m$, then (under mild regularity conditions) for an unbiased estimator $\hat{\theta}(X)$ the inequality of Cramer-Rao indicates that

$$\boldsymbol{Var}(\hat{\theta}(X)) \geq I^{-1} \text{ where } I = \boldsymbol{E}\frac{\partial}{\partial\theta}\log p_\theta(X)\frac{\partial}{\partial\theta}' \log p_\theta(X).$$

The term I on the right is the Fisher information matrix w.r.t. the parameter θ. Under the same mild conditions, for a stochastic process $X_1, X_2, \cdots$ of i.i.d. variables with probability distribution p_θ, the sequence of Maximum Likelihood Estimators

$(\hat{\theta}_T(X_1, \cdots, X_T))_{T\in\mathbf{N}}$ is known to be (first order) efficient, i.e. the limiting distribution of $(\sqrt{T}(\hat{\theta}_T - \theta))_{T\in\mathbf{N}}$ is normal with variance equal to the inverse of the Fisher information matrix I. For a time-invariant autoregressive process $(y_t)_{t\in\mathbf{N}}$ of order n, Mann and Wald (1943) proved that the sequence of maximum likelihood estimates $\hat{a}_T$ of the coefficients $a = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$ based on T observations, is first order efficient, i.e. the limiting distribution of $(\sqrt{T}(\hat{a}_T - a))_{T\in\mathbf{N}}$ is normal with variance equal to the inverse of the *asymptotic* Fisher information matrix w.r.t. the parameter a. Also the sequence of maximum likelihood estimators of the noise-level σ^2, jointly with those of the coefficients, is efficient in this sense; see for instance Hannan and Deistler (1988). The (asymptotic) Fisher information matrix can be used to define a Riemannian geometry on the parameter set. This is the principle of Amari's differential geometrical methods in statistics (1990). He studied higher order efficiency of estimators for parameters of exponential families by exploiting the geometry of the parameter space. He also studied the geometry given by the asymptotic Fisher matrix on the parameter set U_n of stationary time-invariant autoregressive processes of order n; a geometry, based on another connection than the Levi-Civita (1987). We shall study the geometry of U_n, based on the Levi-Civita-connection, in chapter five. This section serves as a reminder for the expression of the asymptotic Fisher matrix w.r.t. coefficients and noise level (a, σ^2). Chapter five contains many expressions of the asymptotic Fisher information w.r.t. other parameters.

Consider a univariate constant AR(n) coefficient process

$$z_t = a_1 z_{t-1} + \ldots a_n z_{t-n} + \sigma e_t, \tag{2.21}$$

where $(e_t)_{t\in\mathbf{Z}}$ is Gaussian white noise with unit covariance.
We associate with this process its *natural parameter* $\Theta = (a_1, \ldots, a_n, \sigma^2)$, and its *associated polynomial* $L_n(x) = \frac{1}{\sigma}(x^n - a_1 x^{n-1} - \ldots - a_n)$. The AR(n) process is called stationary if and only if all the roots of L_n have modulus less than one.
From now on, we shall suppose that the process (2.21) is Gaussian and stationary.
We also associate with it its *covariance function* η: $\eta_i = E z_t z_{t-i}$, $i \in \mathbf{Z}_+$. It is well known that there exists the following relation between the natural parameter and the covariance function:

$$\begin{pmatrix} \eta_0 & \eta_1 & \cdots & \eta_n \\ \eta_1 & \eta_0 & \cdots & \eta_{n-1} \\ \vdots & \vdots & \vdots & \vdots \\ \eta_n & \eta_{n-1} & \cdots & \eta_0 \end{pmatrix} \begin{pmatrix} 1 \\ -a_1 \\ \vdots \\ -a_n \end{pmatrix} = \begin{pmatrix} \sigma^2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{2.22}$$

Let us denote with $p_{L_n}(z_0, \ldots, z_T)$ the joint probability density function of the $T+1$ variables $z_0, \ldots, z_T$ of the Gaussian stationary AR(n) process with associated polynomial L_n. For clearness of exposition, we give now a derivation of the well known asymptotic Fisher information matrix of a Gaussian stationary AR(n) process w.r.t. its natural parameter which appeared in for instance Box and Jenkins, (1976), p.240.

In order to study the Fisher information $G_\Theta(n,T)$ in the distribution $p_{L_n}(z_0,\ldots,z_T)$, $(T \geq n)$ w.r.t. the natural parameter $\Theta = (a_1,\ldots,a_n,\sigma^2)$, we have to calculate:

$$(G_\Theta(n,T))_{ij} = E\frac{\partial \log p_{L_n}}{\partial \Theta_i}\frac{\partial \log p_{L_n}}{\partial \Theta_j} \tag{2.23}$$

Following Bayes' rule we have:

$$p_{L_n}(z_0,\ldots,z_T) = p_{L_n}(z_n,\ldots,z_T \mid z_0,\ldots,z_{n-1})p_{L_n}(z_0,\ldots,z_{n-1}),$$

$$\text{hence, if we denote by: } l_{L_n}(s) := \log(p_{L_n}(z_0,\ldots,z_s)),$$

$$l_{L_n}(T) = -\frac{T+1-n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=n}^{T}(z_t - a_1 z_{t-1} - \ldots - a_n z_{t-n})^2 + l_{L_n}(n-1). \tag{2.24}$$

Differentiating yields:

$$\frac{\partial}{\partial a_i}l_{L_n}(T) = \sum_{t=n}^{T}\frac{e_t z_{t-i}}{\sigma} + \frac{\partial}{\partial a_i}l_{L_n}(n-1);$$

$$\frac{\partial}{\partial \sigma^2}l_{L_n}(T) = \frac{1}{2\sigma^2}\sum_{t=n}^{T}(e_t^2 - 1) + \frac{\partial}{\partial \sigma^2}l_{L_n}(n-1).$$

Now, the first and second terms in these equations are not correlated, and by taking covariances we find:

$$G_\Theta(n,T) = (T+1-n) \times H_\Theta^n + G_\Theta(n,n-1), \tag{2.25}$$

where

$$H_\Theta^n = \begin{pmatrix} \frac{\eta_0}{\sigma^2} & \frac{\eta_1}{\sigma^2} & \ldots & \frac{\eta_{n-1}}{\sigma^2} & 0 \\ \frac{\eta_1}{\sigma^2} & \frac{\eta_0}{\sigma^2} & \ldots & \frac{\eta_{n-2}}{\sigma^2} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\eta_{n-1}}{\sigma^2} & \frac{\eta_{n-2}}{\sigma^2} & \ldots & \frac{\eta_0}{\sigma^2} & 0 \\ 0 & 0 & \ldots & 0 & \frac{1}{2\sigma^4} \end{pmatrix} \tag{2.26}$$

H_Θ^n will be called the *asymptotic Fisher information matrix* w.r.t the parameters Θ. Note that it follows from equation (2.22) that the elements of H_Θ^n depend on Θ in a differentiable way.

2.5 The theory of shift invariant inner products

This section serves as a reminder of a theory that we shall use in chapter five to investigate other important parametrizations of the set U_n of univariate stationary time-invariant autoregressive processes of order n, than the natural one. One parametrization that follows from the theory presented below, identifies this set U_n with $(-1,1) \times \cdots \times (-1,1) \times \mathbb{R}^+$. If one uses the natural parametrization the situation will be far more complicated. The theory also has implications for identification algorithms: Levinson's algorithm can be based on it. We shall even use this theory incidently in

chapter six to discuss asymptotic behaviour of the Kalman Filter for a particular dynamical stochastic system with time-invariant system matrices. All the results in this section are well known, and can be found in e.g. H.J. Landau (1987).
First, we introduce some notation and terminology:

- A polynomial with real coefficients will be called *stable* if all its roots have modulus less than one;
- The linear space of polynomials with real coefficients and degree $\leq n$ will be denoted by Π_n. This space will be identified with $\mathbb{R}^{n+1}$ by means of the trivial parametrization:

$$\mathbb{R}^{n+1} \to \Pi_n; (b_0, \ldots, b_n) \mapsto b_0 x^n + b_1 x^{n-1} + \ldots + b_n;$$

- The subset of Π_n , consisting of the stable polynomials of degree $\leq n$ with positive highest coefficient (i.e. $b_0 > 0$), will be denoted by U_n. Remark: U_n is open and connected;
- P_n is the map $\Pi_n \to \Pi_n$, defined by $b(x) \mapsto x^n b(\frac{1}{x})$. We shall refer to this map as the "sip"transform. Considered as a linear transformation of $\mathcal{R}^{n+1}$, P_n has the matrix:

$$P_n = \begin{pmatrix} 0 & 0 & \ldots & 0 & 1 \\ 0 & 0 & \ldots & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \ldots & 0 & 0 \\ 1 & 0 & \ldots & 0 & 0 \end{pmatrix}; \; P_n \begin{pmatrix} b_0 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} b_n \\ \vdots \\ b_0 \end{pmatrix}; \tag{2.27}$$

- Throughout this section, we adopt the following notation:
  Let L_n be a polynomial in Π_n, then we denote its sip transform by $E_n = P_n(L_n)$.
  Note that in that case $P_n(xL_{n-1}(x)) = E_{n-1}(x)$.

Next, we present the main results that we use from the theory of shift invariant inner products on Π_n.

An inner product $\langle , \rangle$ on Π_n is called *shift invariant*

- if $\langle a(x), b(x) \rangle = \langle xa(x), xb(x) \rangle$ for all polynomials $a, b \in \Pi_{n-1}$
  or, equivalently;
- if the sip transforms P_k are isometric transformations of Π_k for all $k \leq n$
  or, equivalently;
- if the positive definite $n+1$-dimensional square matrix $H = (\langle x^i, x^j \rangle)_{0 \leq i,j \leq n}$
  is Toeplitz, i.e., if it has the form:

$$H = \begin{pmatrix} h_0 & h_1 & \ldots & h_n \\ h_1 & h_0 & \ldots & h_{n-1} \\ \vdots & \vdots & \vdots & \vdots \\ h_n & h_{n-1} & \ldots & h_0 \end{pmatrix} \tag{2.28}$$

Let a shift invariant inner product be defined on Π_n, and let $H = (\langle x^i, x^j \rangle)_{0 \le i,j \le n}$ be of the form (2.28). Then the equation:

$$\begin{pmatrix} h_0 & h_1 & \dots & h_n \\ h_1 & h_0 & \dots & h_{n-1} \\ \vdots & \vdots & \vdots & \vdots \\ h_n & h_{n-1} & \dots & h_0 \end{pmatrix} \begin{pmatrix} 1 \\ -a_1 \\ \vdots \\ -a_n \end{pmatrix} = \begin{pmatrix} \sigma^2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{2.29}$$

has a unique solution $(a_1, \dots, a_n, \sigma^2)$ with $\sigma > 0$.
The inner product defines a polynomial L_n of degree n by:

$$L_n(x) = \frac{1}{\sigma}(x^n - a_1 x^{n-1} - \dots - a_n).$$

Let $E_n = P_n(L_n)$ be its sip transform. Then, it is not very difficult to deduce that L_n and E_n have the following properties:

1. $\langle Q(x), L_n(x) \rangle = 0$ for all polynomials $Q \in \Pi_{n-1}$;

2. if $L_n(x) = (x - \phi)P(x)$ then $\langle L_n, L_n \rangle = (1 - \phi^2)\langle P, P \rangle$, hence L_n is stable: $L_n \in U_n$;

3. E_n is the unique element in Π_n, such that $\langle Q(x), E_n(x) \rangle = \frac{Q(0)}{E_n(0)}$ for all $Q \in \Pi_n$. The polynomial $E_n(0)E_n = \frac{1}{\sigma}E_n$ is called the "zero evaluator" w.r.t. this inner product;

4. $\langle L_n, L_n \rangle = \langle E_n, E_n \rangle = 1$.

On the other hand, let L_n be an arbitrary stable polynomial of degree n with positive highest coefficient. $L_n \in U_n$. L_n defines a shift invariant inner product on Π_n in the following way:

$$\langle a(x), b(x) \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{a(e^{i\phi})b(e^{-i\phi})}{E_n(e^{i\phi})E_n(e^{-i\phi})} d\phi \tag{2.30}$$

As $E_n = P_n(L_n)$ is anti-stable, $\frac{Q(z)}{E_n(z)}$ is holomorphic on the unit disk, and, accordingly, also w.r.t. this inner product $\langle Q, E_n \rangle = \frac{Q(0)}{E_n(0)}$. This shows that there exists a one-to-one relationship between shift invariant inner products on Π_n, and elements of U_n. From this one-to-one relationship, it may also be concluded that, for given $L_n(x) = \frac{1}{\sigma}(x^n - a_1 x^{n-1} - \dots - a_n), (\sigma > 0)$, equation (2.29) has a positive definite

solution $H = \begin{pmatrix} h_0 & h_1 & \dots & h_n \\ h_1 & h_0 & \dots & h_{n-1} \\ \vdots & \vdots & \vdots & \vdots \\ h_n & h_{n-1} & \dots & h_0 \end{pmatrix}$ if and only if $L_n \in U_n$.

We shall call the unique shift invariant inner product, which corresponds to L_n in the way described above, the *Schur inner product induced by* L_n.

Finally, we summarize an important well known result that is the basis of the lattice and Levinson's algorithms.

We shall call it the *proposition of the last Schur parameter:*

Proposition 2.3 Let $n \in \mathbf{Z}$, $n > 0$.
Let f be the map: $\Pi_{n-1} \times \mathbf{R} \to \Pi_n$, defined by:

$$(L_{n-1}, \lambda) \mapsto L_n(x) = xL_{n-1}(x) + \lambda E_{n-1}(x).$$

Then, the restriction of $f : U_{n-1} \times (-1,1) \to U_n$ is a *bijection* . Furthermore:

- $\lambda = \frac{L_n(0)}{E_n(0)}$;
- If $L_n \in U_n$, then the Schur inner product on Π_n, induced by L_n, restricted to Π_{n-1}, is induced by $\sqrt{1-\lambda^2}L_{n-1}$.

If $L_n \in U_n$, then λ will be called *the last Schur parameter of* L_n.

Proof: We shall show that

1. f is a bijection from $\Pi_{n-1} \times \mathbf{R} \setminus \{-1,1\}$ onto $\{L_n \in \Pi_n | L_n(0) \neq \pm E_n(0)\}$;
2. $f(U_{n-1} \times (-1,1)) \subset U_n$;
3. $f^{-1}(U_n) \subset U_{n-1} \times (-1,1)$.

1. Let $L_n = f(L_{n-1}, \lambda)$. Applying the sip transform yields:

$$\left\{ \begin{array}{rcl} L_n(x) & = & xL_{n-1}(x) + \lambda E_{n-1}(x) \\ E_n(x) & = & E_{n-1}(x) + \lambda x L_{n-1}(x) \end{array} \right\} \tag{2.31}$$

   If $|\lambda| \neq 1$, this is equivalent to:

$$\left\{ \begin{array}{rcl} xL_{n-1}(x) & = & \frac{1}{1-\lambda^2} L_n(x) - \frac{\lambda}{1-\lambda^2} E_n(x) \\ E_{n-1}(x) & = & \frac{1}{1-\lambda^2} E_n(x) - \frac{\lambda}{1-\lambda^2} L_n(x) \end{array} \right\} \tag{2.32}$$

   Equations (2.31) show that $E_n(0) = E_{n-1}(0) = \frac{1}{c}$, and $\lambda = \frac{L_n(0)}{E_n(0)}$. The proof of 1 is now immediate.

2. Let $L_{n-1} \in U_{n-1}$ and $|\lambda| < 1$, and let $L_n = f(L_{n-1}, \lambda)$.
   Let $\langle , \rangle$ be the Schur inner product, induced by L_{n-1} on Π_{n-1}. We define an inner product on Π_n by taking L_n perpendicular to Π_{n-1}, and by putting $\langle L_n, L_n \rangle = 1 - \lambda^2$. Then, we have that:

$$\langle xL_{n-1}, xL_{n-1} \rangle = 1 = \langle L_{n-1}, L_{n-1} \rangle,$$

   and

$$\langle xL_{n-1}(x), xQ(x) \rangle = 0 = \langle L_{n-1}, Q \rangle$$

   for all $Q \in \Pi_{n-2}$. From this, it follows easily that the so defined inner product is shift invariant, and is induced by $\frac{1}{\sqrt{(1-\lambda^2)}} L_n$, so $L_n \in U_n$.

3. Let $L_n = f(L_{n-1}, \lambda) \in U_n$.
   Let $\langle , \rangle$ be the Schur inner product, induced by L_n on Π_n.
   As $\lambda = \frac{L_n(0)}{E_n(0)} = \langle L_n, E_n \rangle$, it follows that $|\lambda| < 1$ from Schwartz' inequality, property 4, and the fact that $L_n \neq E_n$ (since $E_n \notin U_n$ as $n > 0$).
   Now, equations (2.32) and properties 1 and 3 show that
   for all $Q \in \Pi_{n-1}$, $\langle Q, E_{n-1} \rangle = \frac{Q(0)}{(1-\lambda^2)E_n(0)}$. Hence, the restriction of the inner product to Π_{n-1} is the Schur inner product, induced by $\sqrt{1-\lambda^2}L_{n-1}$, from which we see that $L_{n-1} \in U_{n-1}$.

2.6 Kalman Filtering and Smoothing; The Extended Kalman Filter

We assume that the reader is familiar with the principles of Kalman Filtering, (fixed interval) Smoothing and the usual way to extend these identification techniques in order to estimate states in non-linear dynamical stochastic systems. Appropriate references are e.g. P.C.Young (1984), Jazwinski (1970) and Anderson and Moore (1979).

2.7 Arzela-Ascoli; Ornstein-Uhlenbeck

In this section, we gathered the definitions of equicontinuity, the equiLipschitz property for collections of functions, and the theorem of Arzela-Ascoli. We included these notions only as a reminder. A good reference for these topics is A.N.Kolmogorov and S.V.Fomin (1970), pp. 102-107. We shall use these notions in chapter four, where we additionally introduce the concept of "equislowly" varying coefficients of a sequence of time-varying autoregressive processes for developing a sound asymptotic theory. In the same chapter we also use the concept of the well-known Ornstein-Uhlenbeck Velocity process. We shall consider some basic aspects of this continuous-time stochastic process.

Definition. Let X and M be metrical spaces with distances $\boldsymbol{d}_X$ and $\boldsymbol{d}_M$ respectively. A collection F of functions: $X \to M$ is called *equicontinuous* if for every $a \in X$ and every $\varepsilon > 0$ there exists a neighbourhood $U_{a,\varepsilon}$ of a, such that

$$\boldsymbol{d}_M(f(x), f(a)) < \varepsilon \text{ for all } x \in U_{a,\varepsilon} \text{ and all } f \in F.$$

The collection F is called *equiLipschitz,* if there exist positive numbers η and C, such that

$$\text{if } \boldsymbol{d}_X(x,a) < \eta \text{ then } \boldsymbol{d}_M(f(x), f(a)) < C\boldsymbol{d}_X(x,a) \text{ for all } a, x \in X \text{ and all } f \in F.$$

Any collection of functions $X \to M$ that is equiLipschitz, is also equicontinuous. Moreover, all functions from such an equiLipschitz collection are uniformly continuous on X. Furthermore, they can be extended in a unique way to continuous functions on $\bar{X}$, the metric completion of X. The resulting collection of functions $\bar{X} \to M$ is again

equiLipschitz.

Theorem of Arzela-Ascoli. Let X and M be metrical spaces; X compact. Let F be an equicontinuous collection of functions from X to M. Suppose that for any $a \in X$ the set $\{f(a) \mid f \in F\}$ has compact closure in M. Then, every sequence of functions in F has a subsequence which converges uniformly on X to a continuous function $X \to M$.

Now, we give a reminder on the Ornstein-Uhlenbeck Velocity (OUV) process. For a definition and existence proof of an OUV process, we refer to Doob (1942). Continuity of the sample paths is already stated in that article, but this property can also be proved using a theorem of Kolmogorov (see e.g. Gard, (1988)).

Doob (1942) has proved that there exist univariate stochastic processes $(u_t)_{t\in\mathbf{R}}$, defined on the whole real axis, which satisfy the following properties:

- $(u_t)_{t\in\mathbf{R}}$ is stationary;
- if $s = \max\{s_1, \cdots, s_k\}$ and $t > s$ then $p(u_t|u_{s_1}, \cdots, u_{s_k}) = p(u_t|u_s)$ (Markov-property);
- $\forall s, t \in \mathbf{R}, s \neq t$, we have $\begin{pmatrix} u_s \\ u_t \end{pmatrix} \stackrel{d}{=} \mathcal{N}(0, V)$ with $\det(V) \neq 0$.

The probability densities of such a process are completely determined by two positive numbers β and τ in the sense that

$$\boldsymbol{E}u_t u_s = \tau^2 e^{-\beta|t-s|} \tag{2.33}$$

We shall not discuss the possibility that $\beta = \infty$, i.e. that $(u_t)_{t\in\mathbf{R}}$ is "white noise".
As $\boldsymbol{E}(u_{s+h} - u_s)^2 = 2\tau^2(1 - e^{-\beta|h|})$, the process is mean-square-continuous. One can use a theorem of Kolmogorov (Gard, 1988) to prove continuity of the sample paths $t \mapsto u_t(\omega)$, $(\omega \in \Omega)$ P-almost sure: The theorem states that, if there exists h_0, such that

$$\boldsymbol{E}|u_{t+h} - u_t|^\alpha \leq C|h|^{1+\delta} \; \forall h, |h| \leq h_0$$

for some positive numbers α, δ, C, then the sample paths are continuous P-a.s.. Since $u_{t+h} - u_t$ is Gaussian, we have:

$$\boldsymbol{E}|u_{t+h} - u_t|^4 = 12\tau^4(1 - e^{-\beta|h|})^2,$$

hence, the condition of Kolmogorov is surely satisfied.
If one integrates this process, the result is called an Ornstein-Uhlenbeck process. The process u itself is usually called an Ornstein-Uhlenbeck Velocity process, characterized by the positive constants β and τ: we shall refer to it as an OUV(β, τ)-process.
Due to the autocovariance structure given in (2.33), (or the markov-property) some authors call the Ornstein-Uhlenbeck Velocity process a *continuous autoregressive process*

of order one (CAR(1); see Priestley (1981, section 3.7), Chan and Tong (1987), He and Wang (1989)). It is easily seen that, for all $T \in \mathbb{R}^+$, the discrete-time process $(x_t)_{t\in\mathbf{Z}}$,

$$x_t = u(\frac{t}{T}) \quad (t \in \mathbf{Z})$$

is a time-invariant, stationary autoregressive process of order one, satisfying the equation

$$x_{t+1} = e^{-\frac{\beta}{T}} x_t + \sqrt{1 - e^{-\frac{2\beta}{T}}}\tau\lambda_t, \tag{2.34}$$

where $(\lambda_t)_{t\in\mathbf{Z}}$ is Gaussian white noise with variance one.

Chapter 3

Polygonal processes on a Riemannian manifold

In this chapter, we present a model for the coefficients of a univariate AR(n) process $(y_t)_{t\geq 0}$ if these are time-varying.
Consider such a process:

$$y_t = a_{t,1}y_{t-1} + \ldots + a_{t,n}y_{t-n} + \sigma_t^2 e_t, \tag{3.1}$$

where

- $\sigma_t, a_{t,i}(1 \leq i \leq n; t \geq n)$ are given numbers, such that $\sigma_t > 0$ for all $t \in \mathbf{Z}_+$
- $(e_t)_{t\in\mathbf{Z}_+}$ i.i.d $\mathcal{N}(0,1)$, and independent of $y_{-1}, \ldots, y_{-n}$
- $(y_{-1}, \ldots, y_{-n})'$ has zero expectation, and $\boldsymbol{Var}(y_{-1}, \ldots, y_{-n})$ is positive definite.

Then, one can show by induction that $\boldsymbol{Var}(y_t, \ldots, y_{-n})$ is positive definite for all $t \geq n$, and that the numbers $\sigma_t, a_{t,i}(1 \leq i \leq n; t \geq n)$ are uniquely determined by the process, because the equation:

$$\boldsymbol{Var}(y_t, \ldots, y_{-n}) \begin{pmatrix} 1 \\ -a_{t,1} \\ \vdots \\ -a_{t,n} \\ -a_{t,n+1} \\ \vdots \\ -a_{t,t} \end{pmatrix} = \begin{pmatrix} \sigma_t^2 \\ 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{pmatrix} \tag{3.2}$$

has a unique solution $\sigma_t > 0, a_{t,1}, \ldots, a_{t,t}$. (Necessarily, $a_{t,n+1} = \ldots = a_{t,t} = 0$.). Hence, the following definition makes sense if $t \in \mathbf{Z}_+$:
The *instantaneous associated polynomial* denoted as $L_{n,t}$ or q_t of the process (3.1) at time t is the polynomial:

$$q_t = L_{n,t} = \frac{1}{\sigma_t}(x^n - a_{t,1}x^{n-1} \ldots - a_{t,n}) \tag{3.3}$$

The numbers $a_{t,1}, \ldots, a_{t,n}, \sigma_t$ can be the coefficients of a univariate Gaussian stationary constant AR(n) coefficient system if and only if the instantaneous associated polynomial $L_{n,t}$ is stable.
It is possible to view the process (3.1) as an evolution of Gaussian stationary constant AR(n) coefficient systems into each other as long as the instantaneous associated polynomials are stable. We will study the process from this point of view in the next chapter.
Note that as long as the stability condition is satisfied, we can consider the process (3.1) as *a walk of the instantaneous associated polynomials* $(q_t)_{t\in\mathbf{Z}_+}$ *through the set* U_n *of all stable polynomials of degree n with positive highest coefficient.*
We would like to model this walk stochastically, taking a relevant geometrical structure of U_n (or another coefficient space) into account.
Among those authors who like a stochastic model for the coefficients, it has been customary to model the coefficients $(q_t)_{t\in\mathbf{Z}_+}$ as a (smoothly integrated) random walk

$$\begin{aligned} q_{t+1} &= q_t + d_t \\ d_t &= ad_{t-1} + v_t \quad (a \in [0,1)\,) \end{aligned}$$

where v_t is independent of $q_0, d_0, \cdots, d_{t-1}, q_t$. Connecting the coefficients q_t, one gets a polygon with edges equal to the directions d_t. The length of the term v_t can be interpreted as the amount that the direction d_t is varying from its expected value ad_{t-1}. This length is independent of the past; it does not depend on where q_t is in the coefficient space. But the terms "polygon" and "length" are used in the Euclidean sense. If we take another type of geometry, these terms will get another meaning. Now, any Riemannian manifold structure on the coefficient space defines a geometry. So, it seems reasonable to generalize the notions "polygonal" and "independent directional variation" for a stochastic process on the coefficient space that is in agreement with its geometry. For instance, in the usual model, the edge of the polygon between q_t and q_{t+1} is a piece of a *straight line* in the direction d_t. We shall replace this edge by a piece of the *geodesic* from q_t in the direction d_t. Furthermore, the term "independent directional variation" will mean that the directional variation v_t will have a *Riemannian* length that is independent of the past.
The terms *polygonal process* and *independent directional variation* (and some technicalities) will be introduced in the sections (3.1), (3.2), (3.3), and (3.4). In section (3.5) we present methods to describe these polygonal processes of independent directional variation and their probability distributions. *Direction spaces* and *eigen distributions of a symmetric tensor field* will play an important role. Then we come to the kernel of the chapter. We formulate two models for such polygonal coefficient processes, a Special Model (a generalization of the usual random walk) in section (3.6), and a General Model (a generalization of the smoothly integrated random walk) in section (3.7). We study the linearity of these models in section (3.8), their complexity in section (3.9), and convergence properties of processes, generated according to the Special and General Models in section (3.10).

3.1 Polygonal processes and their running times

Let M be an n-dimensional Riemannian manifold. For any point $q \in M$ and any vector d in the tangent space M_q at q, there exists exactly one geodesic $\gamma_{q,d} : \mathbb{R} \to M$, such that $\gamma_{q,d}(0) = q$ and $\dot{\gamma}_{q,d}(0) = d$. For any point $q \in M$, there exists a neighbourhood V_q of 0 in the tangents space M_q, such that the following map is C^∞, and its restriction to V_q has an inverse which is also C^∞:

$$\begin{aligned} \exp_q : M_q &\to M \\ \exp_q(d) &= \gamma_{q,d}(1). \end{aligned} \tag{3.4}$$

This map is called the *normal map in* q. Its domain consists of all the vectors $d \in M_q$, such that the geodesic starting in q in the direction d has a length longer than 1.
We also form the *tangent bundle* TM of the manifold M:

$$TM = \{(q,d) | q \in M; d \in M_q\} \tag{3.5}$$

With its Borel sets as σ-algebra, the manifold M, and also the tangent bundle TM are measurable spaces. For more details on the topology of M and TM and the normal map, see chapter two.
Let $(\Omega, \mathcal{A}, P)$ be a probability space.
Let $(\mathcal{B}_t)_{t \in \mathbf{Z}, t \geq -1}$, $\mathcal{B}_t \subset \mathcal{A}$ be an increasing sequence of σ-algebras.
We shall consider a special class of stochastic processes on the tangent bundle TM, adapted to this increasing sequence of σ-algebras.

Definition. Let $(q_t, d_t)_{t \in \mathbf{Z}_+}, (q_t, d_t) : \Omega \to TM$, be a stochastic process. We say that it is *a polygonal process adapted to* $(\mathcal{B}_t)_{t \in \mathbf{Z}, t \geq -1}$ if

- (q_t, d_t) is measurable with respect to the σ-algebra $\mathcal{B}_t$ for all $t \in \mathbf{Z}_+$
- $q_0 : \Omega \to M$ is $\mathcal{B}_{-1}$-measurable, and
- $q_{t+1} = \exp_{q_t}(d_t)$ for all $t \in \mathbf{Z}_+$ for which the right hand side is well defined. □

With such a process we associate a special, discretely valued, stochastic variable ν:

$$\nu = \inf\{t \in \mathbf{Z}_+ | \ \exp_{q_t}(d_t) \text{ is not well defined } \} \tag{3.6}$$

It is well known that, if M is complete, $\exp_q$ is defined on the whole tangent space M_q for all $q \in M$, so in that case $\nu = \infty$. We shall often work, however, with manifolds which are not complete. The next lemma 3.1 shows that ν is a stopping time.

Lemma 3.1 The stochastic variable ν is a stopping time with respect to the sequence $(\mathcal{B}_t)_{t \in \mathbf{Z}, t \geq -1}$, i.e. the set $\{\omega \in \Omega | \ \nu(\omega) > t\}$ is an element of the σ-algebra $\mathcal{B}_t$ for all $t \in \mathbf{Z}_+$.

Proof: We use a fact from the theory of ordinary differential equations:
Let $p \in \mathbb{N}$ and $V \subset \mathbb{R}^{p+1}$ an open neighbourhood of $(0, w)$ for some $w \in \mathbb{R}^p$. A solution $x : I \to \mathbb{R}^p$ of an ordinary differential equation

$$\dot{x}(t) = f(t, x), \text{ where } f : V \to \mathbb{R}^p \text{ continuous and locally Lipschitz in } x,$$

is called maximal if its domain $I \subset \mathbb{R}$ can't be extended. Let the unique maximal solution $x(.,x_0)$ of this differential equation with starting value $x(0,x_0) = x_0 \in \mathbb{R}^p$ be defined on I_{x_0}. Then, it is a well known fact that the set

$$\{(t,x_0) \in V |\ t \in I_{x_0}\}$$

is open in $\mathbb{R}^{p+1}$. It is no problem to extend the validity of this fact for ordinary differential equations on $\mathbb{R}^p$ to ordinary differential equations on a differential manifold N; just use coordinate neighbourhoods. Consider the maximal solutions $\gamma(t,q,d)$ of the differential equation on TM

$$\nabla_{\dot\gamma(t,q,d)}\dot\gamma(t,q,d) = 0\ ;\ \gamma(0,q,d) = q\ ;\ \dot\gamma(0,q,d) = d.$$

These are the geodesics starting in $q \in M$ in the direction $d \in M_q$. From the above mentioned fact from the theory of ordinary differential equations, we see that the set

$$W = \{(t,q,d) \in \mathbb{R} \times TM | \gamma(t,q,d) \text{ is well defined }\}$$

is open in $\mathbb{R} \times TM$.
We have to show that $\{\nu > t\}$[1] is element of $\mathcal{B}_t$. Well, note that

$$\{\nu > t\} = \bigcap_{s \le t} \{(1, q_s, d_s) \in W\},$$

which is surely measurable w.r.t. $\mathcal{B}_t$. Hence, ν is a stopping time. □

The variable ν will be called the *running time* of a polygonal process. If $(q_t, d_t)_{t\in\mathbb{Z}_+}$ is a polygonal process, adapted to $(\mathcal{B}_t)_{t\in\mathbb{Z}_+}$, and ρ is a stopping time smaller than ν, then the set $\{\rho > t-1\}$ is contained in the set $\{\nu > t-1\}$. Hence, on this set q_t is equal to $q_t = \exp_{q_{t-1}}(d_{t-1})$, so on this set q_t is not only $\mathcal{B}_t$-measurable, but also $\mathcal{B}_{t-1}$-measurable. If $(q_t, d_t)_{t\in\mathbb{Z}_+}$ is a polygonal process, and ρ is a stopping time, also the *stopped* process $(\tilde q_t, \tilde d_t)_{t\in\mathbb{Z}_+}$, defined by:

$$(\tilde q_t, \tilde d_t) = (q_{\min(t,\rho)}, d_t 1_{\{\rho > t\}}) \tag{3.7}$$

is a polygonal process, adapted to $(\mathcal{B}_t)_{t\in\mathbb{Z}, t\ge -1}$.

3.2 The Integrability Condition and Never-Evasiveness

In order to facilitate the treatment of polygonal processes, we need to make some rather technical assumptions. Firstly, we often assume that a polygonal process $(q_t, d_t)_{t\in\mathbb{Z}_+}$ satisfies *the Integrability Condition:*

$$\boldsymbol{E}(||d_t||_{q_t} 1_{\{q_t \in K\}}) < \infty \text{ for all } t \in \mathbb{Z}_+ \text{ and compact } K \subset M.$$

[1] By abuse of notation, we shall write $\{\nu > t\}$ instead of $\{\omega \in \Omega | \nu(\omega) > t\}$.

This is just to guarantee the existence of the $\mathcal{B}_{t-1}$-measurable map $(q_t, \boldsymbol{E}^{\mathcal{B}_{t-1}} d_t)$; we refer to chapter two for the definition. We can then define the *directional variation term*

$$v_t = d_t - \boldsymbol{E}^{\mathcal{B}_{t-1}} d_t.$$

Secondly, we shall frequently suppose that a polygonal process is *never-evasive*, i.e. at any time t, and for any open subset U of M, there is a positive probability of finding the process element q_t in U. In the next lemma 3.2, we give an easily verifiable criterion to establish never-evasiveness for a polygonal process.

Lemma 3.2 Let $(q_t, d_t)_{t\in\mathbf{Z}_+}$ be a polygonal process, adapted to the increasing sequence of σ-algebras $(\mathcal{B}_t)_{t\in\mathbf{Z}, t\geq -1}$ and ν its running time. Suppose that the process satisfies the *Never-Evasiveness Criterion:*

$$P(q_0 \in U \text{ and } \sum_{s=0}^{t} ||d_s||_{q_s} < \varepsilon) > 0 \text{ for all open sets } U \subset M \text{ and all } \varepsilon > 0\ ;\ t \in \mathbf{Z}_+. \tag{3.8}$$

Then, the process is never-evasive, i.e.:

$$P(\nu > t-1 \text{ and } q_t \in U) > 0 \text{ for all open sets } U \subset M \text{ and } t \in \mathbf{Z}_+. \tag{3.9}$$

Proof: Let U be an open set in M. Choose $p \in U$ arbitrarily. There exists a ball $B_{r_p}(p) = \{q \in M |\ d(p,q) < r_p\}$, such that any two points in this ball can be connected by a geodesic within this ball, having a length equal to the distance of these two points[2]. If $q \in B_{r_p}(p)$ with $d(p,q) < \eta$, then $\exp_q$ is defined on a ball V of radius $r_p - \eta$ with center 0 in the tangent space of M_q, and $d(q, \exp_q(v)) = ||v||$ for all $v \in V$. Now, it is easy to see that:

$$\{q_0 \in B_{\frac{1}{2}r_p}(p) \text{ and } \sum_{s=0}^{t} ||d_s||_{q_s} < \frac{1}{2} r_p\} \ \subset\ B_{r_p}(p) \text{ so} \tag{3.10}$$

$$\{q_0 \in B_{\frac{1}{2}r_p}(p) \text{ and } \sum_{s=0}^{t} ||d_s||_{q_s} < \frac{1}{2} r_p\} \ \subset\ \{\nu > t-1 \text{ and } q_t \in U\}. \tag{3.11}$$

This proves the lemma 3.2. □

3.3 Independent Directional Variation

Definition. Let be given a polygonal process $(q_t, d_t)_{t\in\mathbf{Z}_+}$, adapted to the increasing sequence of σ-algebras $(\mathcal{B}_t)_{t\in\mathbf{Z}, t\geq -1}$, satisfying the Integrability Condition. Let be given a stopping time μ, which is almost sure smaller than or equal to the running time ν of the polygonal process.
We shall say that this process is *of independent directional variation until* μ if [3]

$$(q_t, d_t) \ = \ (q_t, \boldsymbol{E}^{\mathcal{B}_{t-1}}(d_t) + v_t) \text{ and} \tag{3.12}$$

$$||v_t||_{q_t} \ \text{ is } \ \text{independent of } \mathcal{B}_{t-1} \text{ on the event } \{\mu > t-1\}. \ \square \tag{3.13}$$

[2] "Normal property", see chapter two, or Helgason, Theorem I.9.9.

[3] The definition of "conditional expectation" $(q, \boldsymbol{E}^{\mathcal{B}}(d))$ of a stochastic variable (q, d) with values in the tangent bundle TM is given in chapter two.

If a stopping time ρ is smaller than μ a.s., then a polygonal process of independent directional variation until μ is also of independent directional variation until ρ. If a polygonal process is of independent directional variation until μ, then also the process stopped at μ is of independent directional variation until μ.

We explain the idea behind this definition of independent directional variation .
As shown in Helgason, Proposition I.9.10, for any point $p \in M$, there is a ball $B_{r_p}(p) \subset M$ around p with radius r_p, such that for any points a, b in $B_{r_p}(p)$ there exist uniquely determined vectors $A, B \in M_p$ with lengths shorter than r_p, such that

$$\exp_p(A) = a \text{ and } \exp_p(B) = b$$

and we have:

$$\lim_{(a,b)\to(p,p)} \frac{||A - B||_p}{\boldsymbol{d}(a,b)} = 1.$$

This means that if d_t and $\boldsymbol{E}^{\mathcal{B}_{t-1}} d_t$ are small enough,

$$\boldsymbol{d}(\exp_{q_t}(\boldsymbol{E}^{\mathcal{B}_{t-1}} d_t), q_{t+1}) = \boldsymbol{d}(\exp_{q_t}(\boldsymbol{E}^{\mathcal{B}_{t-1}} d_t), \exp_{q_t}(d_t)) \approx$$

$$\approx ||\boldsymbol{E}^{\mathcal{B}_{t-1}} d_t - d_t||_{q_t} = ||v_t||_{q_t}. \tag{3.14}$$

Now, $\exp_{q_t}(\boldsymbol{E}^{\mathcal{B}_{t-1}} d_t)$ is the position you would expect for the new point q_{t+1}, given the past $\mathcal{B}_{t-1}$. Equation (3.14) shows that the uncertainty in this expected position is approximately equal to $||v_t||_{q_t}$, which does not depend on where the expected position is in M.

3.4 A simple criterion for independent directional variation

We are mainly interested in polygonal processes that are not only never-evasive and satisfy the Integrability Condition, but also satisfy a third condition.
The polygonal process (q_t, d_t), adapted to the increasing sequence of σ-algebras $(\mathcal{B}_t)_{t\in\mathbf{Z}, t\geq -1}$, and satisfying the Integrability Condition, is said to have *a directional variation term that is properly representable in vector fields* if we can write:

$$\begin{aligned} d_t &= \boldsymbol{E}^{\mathcal{B}_{t-1}} d_t + v_t \text{ where} \\ v_t &= \sum_{i=1}^{m} \lambda_{t,i} X_{t,i}(q_t) \qquad (\lambda_t \in \mathbb{R}^m) \ \ (t \in \mathbf{Z}_+) \end{aligned}$$

for some C^∞ vector fields $X_{t,i}$, such that, on the event $\{\nu > t-1\}$, λ_t is independent of $\mathcal{B}_{t-1}$, and has a continuous probability density that is positive on some open set $U \subset \mathbb{R}^m$ around 0.
The set of vector fields $X_{t,1}, \cdots, X_{t,m}$ will be called a set of vector fields *of proper variational representation.* If λ_t is, moreover, Gaussian then we say that the directional

variation term v_t is *normally* representable in the vector fields $X_{t,1}, \cdots, X_{t,m}$. We call $\{X_{t,1}, \cdots, X_{t,m}\}$ a set of vector fields *of normal variational representation.*

We shall discuss a simple criterion for a process to be of independent directional variation. In order to establish it, we need a general lemma.

Lemma 3.3 If a stochastic vector $\lambda : \Omega \to \mathbb{R}^m$ has a probability density which is continuous and positive on some open set $U \subset \mathbb{R}^m$ around 0, then for any vector $x \in \mathbb{R}^m$ we can find a sequence $(B_n)_{n\in\mathbb{N}}$ of measurable sets in $\mathbb{R}^m$ and a positive number c, such that

$$\lim_{n\to\infty} \frac{\boldsymbol{E}\lambda\lambda' 1_{(\lambda\in B_n)}}{P(\lambda \in B_n)} = cxx'. \tag{3.15}$$

Proof: There exists a positive number μ, such that $\mu x \in U$ and an invertible linear map A, such that $Ax = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in AU$ and AU contains the set $[-\epsilon, 1+\epsilon]^m \subset AU$ for some positive number ϵ. The stochastic vector $A\lambda$ has a continuous and positive probability density f on AU. Define the set $B_n = [0,1] \times [-\frac{1}{n}, \frac{1}{n}]^{m-1}$ for $n > \frac{1}{\epsilon}$. For simplicity of notation, we suppose that $m = 2$. Note that for any continuous function $g : AU \to \mathbb{R}$, we have that:

$$\lim_{n\to\infty} \frac{\int_{-\frac{1}{n}}^{\frac{1}{n}} \int_0^1 g(z,y)f(z,y)dzdy}{\frac{2}{n}} = \int_0^1 g(z,0)f(z,0)dz. \tag{3.16}$$

From this, it follows easily that:

$$\lim_{n\to\infty} \frac{\boldsymbol{E}(A\lambda\lambda' A' 1_{(A\lambda\in B_n)})}{P(A\lambda \in B_n)} = \int_0^1 \begin{pmatrix} z^2 & 0 \\ 0 & 0 \end{pmatrix} \frac{f(z,0)}{\int_0^1 f(u,0)du} dz. \tag{3.17}$$

The statement of the lemma 3.3 is now immediate. □

Proposition 3.4 Let $(q_t, d_t)_{t\in\mathbb{Z}_+}$ be a polygonal process, adapted to the increasing sequence of σ-algebras $\mathcal{B}_{t,t\in\mathbb{Z},t\geq -1}$, satisfying the Never-Evasiveness Criterion and the Integrability Condition. Let ν be its running time. Suppose that the process also has a directional variation term v_t which is properly representable in vector fields:

$$v_t = \sum_{i=1}^m \lambda_{t,i} X_{t,i}(q_t). \tag{3.18}$$

Then, the process is of independent directional variation until ν if and only if the vector fields $X_{t,i}$ of proper variational representation have the same inner products everywhere on the manifold, i.e.

$$q \mapsto \langle X_{t,i}(q), X_{t,j}(q)\rangle_q \text{ is constant on } M \tag{3.19}$$

for all $t \in \mathbf{Z}_+, 1 \leq i,j \leq m$.

Proof: If the process is of independent directional variation until ν, then we have for all Borel sets B in $\mathbf{R}^m$:

$$\boldsymbol{E}^{\mathcal{B}_{t-1}}(||v_t||^2_{q_t} 1_{\{\lambda_t \in B\}} 1_{\{\nu > t-1\}}) = \boldsymbol{E}(||v_t||^2_{q_t} 1_{\{\lambda_t \in B\}} 1_{\{\nu > t-1\}}) \text{ on the set } \{\nu > t-1\}. \quad (3.20)$$

Now, we have on the set $\{\nu > t-1\}$ that:

$$\boldsymbol{E}^{\mathcal{B}_{t-1}}(||v_t||^2_{q_t} 1_{\{\nu > t-1\}} 1_{\{\lambda_t \in B\}}) = \boldsymbol{E}^{\mathcal{B}_{t-1}}(\lambda_t'(\langle X_{t,i}, X_{t,j}\rangle_{q_t})_{1 \leq i,j \leq m} \lambda_t 1_{\{\lambda_t \in B\}} 1_{\{\nu > t-1\}}) =$$

$$= \mathrm{tr}((\langle X_{t,i}, X_{t,j}\rangle_{q_t})_{1 \leq i,j \leq m} 1_{\{\nu > t-1\}} \boldsymbol{E}(\lambda_t \lambda_t' 1_{\{\lambda_t \in B\}} 1_{\{\nu > t-1\}}))$$

because $X_{t,i}(q_t) 1_{\{\nu > t-1\}}$ is measurable w.r.t. $\mathcal{B}_{t-1}$, and λ_t is independent of $\mathcal{B}_{t-1}$ on the set $\{\nu > t-1\}$. Using lemma 3.3, we see that

$$\mathrm{tr}((\langle X_{t,i}, X_{t,j}\rangle_{q_t})_{1 \leq i,j \leq m} 1_{\{\nu > t-1\}} c x x') = c x' (\langle X_{t,i}, X_{t,j}\rangle_{q_t})_{1 \leq i,j \leq m} 1_{\{\nu > t-1\}} x$$

is constant on the set $\{\nu > t-1\}$ for all $x \in \mathbf{R}^m$. This implies that $(\langle X_{t,i}(q_t), X_{t,j}(q_t)\rangle_{q_t})_{1 \leq i,j \leq m}$ must be constant on the set $\{\nu > t-1\} \subset \Omega$. By the Never-Evasiveness Criterion the process is never-evasive: $P(\nu > t-1 \text{ and } q_t \in U) > 0$ for all open sets $U \subset M$. As the vector fields $X_{t,i}$ are continuous and M is connected, the functions $q \mapsto \langle X_{t,i}(q), X_{t,j}(q)\rangle_q$ must be constant on M for all $t \in \mathbf{Z}_+, 1 \leq i,j \leq m$.

If on the other hand the functions $q \mapsto \langle X_{t,i}(q), X_{t,j}(q)\rangle_q = c_{tij}$ are constant on M for all $t \in \mathbf{Z}_+, 1 \leq i,j \leq m$, then:

$$||v_t||_{q_t} = \sqrt{\sum_{i,j=1}^{m} \lambda_{ti} \lambda_{tj} c_{tij}}.$$

This is independent of $\mathcal{B}_{t-1}$, on the event $\nu > t-1$, because it is a function of λ_t. □

Remark 3.5 The proof demonstrates that most conditions in the proposition, especially the Never-Evasiveness Criterion, are only needed to show *the necessity* of the vector fields $X_{t,i}$ to have inner products constant on the manifold (3.19) if we want the process to be of independent directional variation. Those conditions are not needed to show *the sufficiency*. For instance, if we take $q_0 \in M$ fixed, as we shall do occasionally, then still the process will be of independent directional variation if the vector fields $X_{t,i}$ of proper variational representation have inner products which are constant on the manifold.

3.5 Minimal descriptions of the variation term

If a process $(q_t, d_t)_{t \in \mathbf{Z}_+}$ is polygonal, never- evasive, satisfying the Integrability Condition, and of independent directional variation, with the directional variation term normally representable in vector fields, one might wonder in *how many ways* we can normally represent the directional variation term v_t in vector fields. In other words, we want to describe the collection of all sets of vector fields of normal variational representation.

This description will be given in the next proposition and its corollary. The proposition will also state a justification for the word "independent directional variation", because we shall show that independent directional variation implies that the conditional variance of the direction, given the past, $\boldsymbol{Var}^{\mathcal{B}_{t-1}}(d_t)$, can be given by a symmetric tensor field of type (2,0), independent of the past. For these concepts: *tensor field* and *conditional variance of the process* (q_t, d_t), we refer to chapter two. In addition to the concept of tensor fields, we shall also need the concept of *distributions* in the sense of differential geometry. Also this is explained in chapter two.

Proposition 3.6 Let be given a never-evasive polygonal process $(q_t, d_t)_{t\in\mathbf{Z}_+}$, adapted to a sequence of σ-algebras $(\mathcal{B}_t)_{t\in\mathbf{Z}_+}$, and satisfying the Integrability Condition. Let it be of independent directional variation until the running time ν, and have a directional variation term v_t, which is normally representable in vector fields:

$$\begin{aligned} v_t &= \sum_{i=1}^{n} \lambda_{t,i} X_{t,i}(q_t); \qquad (3.21)\\ \lambda_t &\quad \text{is independent of } \mathcal{B}_{t-1}\\ \lambda_t &\stackrel{d}{=} \mathcal{N}(0, Q_t) \text{ where } Q_t \in \mathbb{R}^{m\times m} \text{ positive definite.} \end{aligned}$$

The distribution (= the $\mathcal{F}_M$-module), generated by the vector fields $X_{t,i}$, will be denoted by $\mathcal{D}_t$. The linear space, generated by the vector fields $X_{t,i}$, will be denoted by $\mathcal{L}_t$. Of course, we have:

$$\mathcal{D}_t(p) = \mathcal{L}_t(p) \subset M_p.$$

Then,

1. For every $t \in \mathbf{Z}_+$, there is an everywhere orthonormal set of C^∞-vector fields of normal variational representation at time t, $E_{t,j}$, $1 \le j \le k_t$, such that these vector fields span $\mathcal{L}_t$.

2. For every $t \in \mathbf{Z}_+$, there exists a symmetric tensor field V_t of type $(2,0)$, such that

   (a) $\boldsymbol{Var}^{\mathcal{B}_{t-1}}(d_t)(X,Y) = V_t(X(q_t), Y(q_t))$ for all C^∞ vector fields $X, Y \in \mathcal{D}M$;

   (b) V_t can be regarded as a symmetric map from the tangent spaces M_p into M_p and the eigenvalues τ_{ti}^2, and their multiplicities m_{ti} of V_t are constant, and non-negative on the manifold; V_t can even be considered as a (symmetric) linear transformation of $\mathcal{L}_t$;

   (c) $\mathcal{D}_t(p) = \mathcal{L}_t(p) = (\text{Ker } V_t)^\perp \subset M_p$ for all $p \in M$.

3. For every $t \in \mathbf{Z}_+$, a set of C^∞ vector fields is of normal variational representation at time t if and only if the set is spanning $\mathcal{L}_t$.

Proof:

1. By the fact that the tangent vectors $X_{t,i}(q)$ have inner products, which are constant on the manifold, we can define an inner product on the space $\mathcal{L}_t$ by defining:

$$\langle X_{t,i}, X_{t,j}\rangle = \langle X_{t,i}(q), X_{t,j}(q)\rangle_q.$$

Now, we can apply the Gram-Schmidt procedure on these vector fields $X_{t,i}$. We obtain an orthonormal set of k_t C^∞ vector fields $E_{t,j}$ spanning $\mathcal{L}_t$, such that

$$X_{t,i}(p) = \sum_{j=1}^{n} C_{t\ ji} E_{t,j}(p) \text{ for all } p \in M.$$

The coordinates of the vectors $X_{t,i}(p)$, with respect to the base $(E_{t,j}(p))_{1\le j\le k_t}$ of $\mathcal{D}_t(p)$, are in column number i of the matrix C_t. Of course, the matrix C_t is constant on the manifold M and surjective. Then, we can write:

$$v_t = \sum_{i=1}^{m} \lambda_{t,i} X_{t,i} = \sum_{i=1}^{m}\sum_{j=1}^{k_t} \lambda_{t,i} C_{tji} E_{t,j} = \sum_{j=1}^{k_t} (C_t\lambda_t)_j E_{t,j}.$$

Let $\eta_t = C_t\lambda_t : \Omega \to \mathbb{R}^{k_t}$. As λ_t is Gaussian and λ_t satisfies (3.21), its variance Q_t is positive definite, and also η_t is independent of $\mathcal{B}_{t-1}$ and Gaussian. Now, $\boldsymbol{Var}(\eta_t) = C_tQ_tC_t'$, which is also positive definite, because C_t has full rank. This shows that the representation of v_t in the $E_{t,j}$ is proper and normal.

2. On any chart $(\tilde{U}_\alpha, x_\alpha)$, we can extend the orthonormal set $E_{t,j}$ $(1 \le j \le k_t)$ to a moving orthonormal base of C^∞-vector fields $E_{t,1}, \cdots, E_{t,k_t}, E_{t,k_t+1}, \cdots, E_{t,n}$ just by applying the Gram-Schmidt procedure on the set $E_{t,1}, \cdots, E_{t,k_t}, \frac{\partial}{\partial x_{\alpha\ 1}}, \cdots, \frac{\partial}{\partial x_{\alpha\ n}}$. The new vector fields $E_{t,j}$, $j \ge k_t + 1$ are all perpendicular to $\mathcal{D}_t$.

   (a) For the vector fields $E_{t,j}$ $(1 \le j \le k_t)$, we have:

$$\langle v_t, E_{t,i}\rangle_{q_t} = \sum_{j=1}^{k_t} \eta_{t,j}\langle E_{t,j}, E_{t,i}\rangle_{q_t} = \eta_{t,i};$$

$$\boldsymbol{Var}^{\mathcal{B}_{t-1}}(d_t)(X,Y) = \boldsymbol{Var}^{\mathcal{B}_{t-1}}(v_t)(X,Y) = \boldsymbol{E}^{\mathcal{B}_{t-1}}(\langle v_t, X(q_t)\rangle_{q_t}\langle v_t, Y(q_t)\rangle_{q_t});$$
$$\boldsymbol{Var}^{\mathcal{B}_{t-1}}(d_t)(E_{t,i}, E_{t,j}) = \boldsymbol{E}^{\mathcal{B}_{t-1}}(\eta_{t,i}\eta_{t,j}) = \boldsymbol{E}(\eta_{t,i}\eta_{t,j}) = (C_tQ_tC_t)_{ij},$$

because η_t is independent of $\mathcal{B}_{t-1}$. If, for any vector field X, $X(q_t)$ is perpendicular to $\mathcal{D}_t(q_t)$ then, for any other vector field Y, we have:

$$\langle v_t, X(q_t)\rangle_{q_t} = 0 \text{ hence } \boldsymbol{Var}^{\mathcal{B}_{t-1}}(d_t)(X,Y) = 0.$$

Now, define the symmetric tensor field $V_t : \mathcal{D}M \times \mathcal{D}M \to \mathcal{F}$ by:

$$V_t(E_{t,i}, E_{t,j}) = (C_tQ_tC_t)_{ij}, \text{ if } 1 \le i,j \le k_t;$$

$$V_t(E_{t,i}, E_{t,j}) = 0, \text{ if } k_t + 1 \le i \le n.$$

By this, it is clear that V_t is well defined as a symmetric tensor field of type $(2,0)$, and that

$$\boldsymbol{Var}^{\mathcal{B}_{t-1}}(d_t)(X,Y) = V_t(X(q_t), Y(q_t)).$$

(b) Any symmetric tensor field of type $(2,0)$, other than the Riemannian inner product, induces a symmetric tensor field of type $(1,1)$, (see chapter two) and hence a symmetric linear transformation of every tangent space M_p. We denote also the symmetric tensor field of type $(1,1)$ by V_t. The matrix of V_t, restricted to $\mathcal{L}(p)$ on the basis $E_{t,j}$ $(1 \leq k_t)$, is $C_t Q_t C_t'$, which is constant on the manifold. This shows that one can consider V_t as a (symmetric) linear transformation of $\mathcal{L}_t$. The eigenvalues of V_t, viewed as a symmetric map from M_p into M_p, are the eigenvalues of the matrix $C_t Q_t C_t'$. Hence they are constant on the manifold, and the multiplicity of the eigenvalue zero is $m_{t0} = n - k_t$. The eigenvalues are non-negative, because $C_t Q_t C_t'$ is positive definite.

(c) On any chart, $\mathrm{Ker}(V_t)(p)$ is spanned by the vectors $E_{t,j}(p)$, $(k_t + 1 \leq j \leq n)$, and $\mathcal{D}_t(p) = \mathcal{L}_t(p)$ is spanned by the vectors $E_{t,j}(p)$, $(1 \leq j \leq k_t)$.

3. Suppose that the directional variation term is normally representable in the vector fields $Y_{t,1}, \cdots, Y_{t,s}$. These vector fields generate a linear space, denoted by $\tilde{\mathcal{L}}_t$. We have to show that $\mathcal{L}_t = \tilde{\mathcal{L}}_t$. Apply part 1. of proposition 3.6: there is an everywhere orthonormal set of vector fields $\tilde{E}_{t,j}$ of normal variational representation at time t, and spanning $\tilde{\mathcal{L}}_t$. Applying part 2c. of this proposition shows that

$$\tilde{\mathcal{L}}_t(p) = \mathrm{Ker}(V_t)^{\perp}, \text{ hence } \dim \tilde{\mathcal{L}}_t = k_t \text{ and}$$

$$\boldsymbol{span}\{\tilde{E}_{t,1}(p), \cdots \tilde{E}_{t,k_t}(p)\} = \boldsymbol{span}\{E_{t,1}(p), \cdots E_{t,k_t}(p)\}.$$

Consequently, for every tangent space M_p there exists an orthogonal matrix $U_t(p)$, such that

$$E_{t,j}(p) = \sum_{l=1}^{k_t} U_{tlj}(p)\tilde{E}_{t,l}(p). \text{ Hence,}$$

$$v_t = \sum_{j=1}^{k_t} \eta_{t,j} E_{t,j}(q_t) = \sum_{j=1}^{k_t}\sum_{l=1}^{k_t} \eta_{t,j} U_{tlj}(q_t)\tilde{E}_{t,l}(q_t) = \sum_{l=1}^{k_t} (U_t\eta_t)_l \tilde{E}_{t,l}(q_t).$$

As v_t is normally representable in the vector fields $\tilde{E}_{t,j}$, we also have:

$$v_t = \sum_{l=1}^{k_t} \mu_{t,l}\tilde{E}_{t,l}(q_t)$$

where μ_t is independent of $\mathcal{B}_{t-1}$ and Gaussian with positive definite variance. Hence, by independence of the $\tilde{E}_{t,l}$,

$$\mu_t = U_t(q_t)\eta_t.$$

Independence of μ_t and η_t of $\mathcal{B}_{t-1}$ implies:

$$\boldsymbol{E}\mu_t\mu_t' 1_{\{\eta_t \in B\}} = \boldsymbol{E}^{\mathcal{B}_{t-1}}\mu_t\mu_t' 1_{\{\eta_t \in B\}} = U_t(q_t)\boldsymbol{E}(\eta_t\eta_t' 1_{\{\eta_t \in B\}})U_t(q_t)'$$

for any Borel set $B \in \mathbb{R}^{k_t}$. So, from lemma 3.3 we conclude that

$$U_t(p)xx'U_t(p)' \quad (p \in M; x \in \mathbb{R}^{k_t})$$

does not depend on p. Then, for all $x \in \mathbb{R}^{k_t}$, and all points $p, q \in M$, we have:

$$U_t(p)xx'U_t(p)' = U_t(q)xx'U_t(q)'.$$

Let $A = U_t(p)'U_t(q)'^{-1}$. Then, $A'xx'A = xx'$ for all $x \in \mathbb{R}^{k_t}$. Consequently, also:

$$\langle Au, xx'Au\rangle = \langle u, xx'u\rangle \text{ for all } x, u \in \mathbb{R}^{k_t}.$$

It follows immediately that $A = I$ or $A = -I$. The manifold M is connected and the orthogonal matrices $U_t(p)$ depend continuously on p. Consequently, $U_t(p) = U_t(q)$ for all points $p, q \in M$, and $\tilde{\mathcal{L}}_t = \mathcal{L}_t$.
Now, the converse. Suppose that $Y_{t,1}, \cdots, Y_{t,s}$ is a set of vector fields spanning $\mathcal{L}_t$. By re-ordering we may suppose that $Y_{t,1}, \cdots, Y_{t,k_t}$ is a base of $\mathcal{L}_t$, and that $Y_{t,k_t+j} = \sum_{l=1}^{k_t} G_{tlj}Y_{t,l}$, and $E_{t,j} = \sum_{l=1}^{k_t} W_{tlj}Y_{t,l}$. The term v_t is normally representable in $E_{t,j}$, hence $v_t = \sum_{j=1}^{k_t} \eta_{tj}E_{t,j}(q_t)$, where $\eta_t \in \mathbb{R}^{k_t}$ is zero mean Gaussian with positive definite covariance. Let $(\rho_{t,1}, \cdots, \rho_{t,s-k_t})'$ also be zero mean Gaussian with positive definite covariance, independent of η_t and independent of $\mathcal{B}_{t-1}$. Then, $\begin{pmatrix} \eta_t \\ \rho_t \end{pmatrix}$ is zero mean Gaussian with positive definite covariance. Now, we can write:

$$v_t = \sum_{j=1}^{k_t} \eta_{t,j}E_{t,j} + \sum_{j=1}^{s-k_t} \rho_{t,j}(Y_{t,k_t+j} - \sum_{l=1}^{k_t} G_{tlj}Y_{t,l}) =$$

$$= \sum_{l=1}^{k_t}(\sum_{j=1}^{k_t} W_{tlj}\eta_{t,j} - \sum_{j=1}^{s-k_t} G_{tlj}\rho_{t,j})Y_{t,l} + \sum_{j=1}^{s-k_t} \rho_{t,j}Y_{t,k_t+j} =$$

$$= \sum_{l=1}^{k_t}(W_t\eta_t - G_t\rho_t)_l Y_{t,l} + \sum_{l=k_t+1}^{s} \rho_{t,l-k_t}Y_{t,l} =$$

$$= \sum_{l=1}^{s} \kappa_{t,l}Y_{t,l} \text{ where } \kappa_t = \begin{pmatrix} W_t & -G_t \\ 0 & I \end{pmatrix}\begin{pmatrix} \eta_t \\ \rho_t \end{pmatrix}.$$

As the matrix $W_t \in \mathbb{R}^{k_t \times k_t}$ is non-singular, κ_t is zero mean Gaussian with positive definite variance and independent of $\mathcal{B}_{t-1}$, showing that the set $Y_{t,1}, \cdots, Y_{t,s}$ is of normal variational representation at time t. □

Corollary 3.7 There is an everywhere orthonormal set of C^∞-vector fields $F_{t,m_{t0}+k}$ spanning $\mathcal{L}_t$, $1 \le k \le k_t$, such that we can write:

$$\begin{aligned} d_t &= \boldsymbol{E}^{\mathcal{B}_{t-1}} d_t + v_t \\ v_t &= \sum_{k=1}^{k_t} \tilde{\lambda}_{t,k}\tau_{t,k}F_{t,m_{t0}+k}, \end{aligned} \tag{3.22}$$

where the process $(\tilde{\lambda}_t)_{t\in\mathbb{Z}_+}$ is Gaussian, such that $\boldsymbol{Var}(\tilde{\lambda}_t) = I_{k_t}$, $\tilde{\lambda}_t$ is independent of $\mathcal{B}_{t-1}$ and

$$0 < \tau_{t,1} \le \cdots \le \tau_{t,k_t}.$$

Proof: V_t, restricted to $\mathrm{Ker}(V_t)^\perp$, has the matrix $C_t Q_t C_t'$ with respect to the orthonormal base $(E_{t,j})_{1\le j\le k_t}$, and this matrix is constant on the manifold. Hence, there exists an orthonormal set of eigenvector fields $(F_{t,m_{t0}+k})_{1\le k\le k_t}$, corresponding to the positive eigenvalues $0 < \tau_{t,1}^2 \le \tau_{k_t}^2$, such that

$$F_{t,m_{t0}+k} = \sum_{j=1}^{k_t} U_{t\ jk} E_{t,j} \text{ for all } 1 \le k \le k_t \tag{3.23}$$

where the k_t-dimensional orthogonal matrix U_t is constant on the manifold. Then,

$$v_t = \sum_{j,k=1}^{k_t} \eta_{t,j} U_{t\ jk} F_{t,m_{t0}+k} = \sum_{k=1}^{n} (U_t \eta_t)_k F_{t,m_{t0}+k},$$

where

$$\boldsymbol{Var}(U_t\eta_t) = U_t C_t Q_t C_t' U_t' = \begin{pmatrix} \tau_{t1}^2 & 0\cdots & 0 & \\ 0 & \tau_{t2}^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \tau_{tk_t}^2 \end{pmatrix}.$$

By defining:

$$\tilde{\lambda}_t = \begin{pmatrix} \tau_{t1}^{-1} & 0\cdots & 0 & \\ 0 & \tau_{t2}^{-1} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \tau_{tk_t}^{-1} \end{pmatrix} U_t \eta_t$$

we get the expression (3.22). □

According to proposition (3.6) and corollary (3.7), the variation term v_t completely determines the symmetric tensor field V_t and the vector space $\mathcal{L}_t$. We shall call $\mathcal{L}_t$ the *direction space* at time t of the polygonal process. The inner products of the vector fields in $\mathcal{L}_t$ are constant on the manifold. However, if we don't want to distinguish between observationally equivalent processes, we sometimes do not need to specify this direction space.

Observational Equivalence On any chart $(\tilde{U}_\alpha, x_\alpha)$, we can extend the orthonormal set of vector fields $F_{t,m_{t0}+k}$, $1 \le k \le k_t$ to an orthonormal base $(F_{t,k})_{1\le k\le n}$, using the Gram-Schmidt procedure. All these vector fields are eigenvectors of the tensor field V_t. The eigenvector fields $F_{t,1+\sum_{j=0}^{i-1} m_{tj}}, \cdots, F_{t,\sum_{j=0}^{i-1} m_{tj}}$, corresponding to the eigenvalue τ_{ti}^2, generate a distribution $\mathcal{D}_{t,\tau_{ti}}$ of constant dimension m_{ti}. The tensor field V_t is completely determined by these distributions and the corresponding eigenvalues. On the other hand, as v_t is normally representable in vector fields, it is easy to verify that the probability distribution of v_t, conditional on the past $\mathcal{B}_{t-1}$, is completely determined by its variance which is given by V_t. If $\boldsymbol{E}^{\mathcal{B}_{t-1}} d_t = 0$ a.s., then this is also true for the conditional probability distribution of d_t. Until the running time ν , these conditional probability distributions together with the probability distribution of $q_0 : \Omega \to M$ determine the probability distribution of $(q_s, d_s)_{0\le s\le t}$, conditional on $\{\nu > t\}$ (Bayes' Rule). Conclusion:

Corollary 3.8 *Suppose that two never-evasive polygonal processes are of independent directional variation until their running times,*
have normally in vector fields representable directional variation terms,
satisfying $E^{\mathcal{B}_{t-1}} d_t = 0$ *a.s.* ,
have the same probability distribution for $q_0 : \Omega \to M$,
and have the same V_t *for all* $t \in \mathbf{Z}_+$.
Then, these two processes have the same probability distribution until the running time ν.

These processes need not be equal: If an eigenvalue of V_t has a multiplicity larger than 1, then, for any orthonormal base of eigenvector fields $(F_{t,j})_{1 \le j \le n}$, we can form an other orthonormal base of eigenvector fields $(G_{t,i})_{1 \le i \le n}$, such that the coordinates of the $G_{t,i}(p)$ w.r.t. the base $(F_{t,j})(p)_{1 \le j \le n}$ are not constant on the manifold. Then, the $G_{t,i}$'s are not linear combinations of the $F_{t,j}$, hence the corresponding processes are not equal according to proposition 3.6 part 3. □

Now, we are able to formulate our first Model for a polygonal process of independent directional variation, which will be used to model the process of the coefficients $(q_t)_{t \in \mathbf{Z}_+}$ of a timevarying AR(n) coefficient process $(y_t)_{t \in \mathbf{Z}_+}$, $\langle q_t, \begin{pmatrix} y_t \\ \vdots \\ y_{t-n} \end{pmatrix} \rangle = e_t$, $e_t \overset{i.i.d.}{=} \mathcal{N}(0,1)$.
Note that the geometry of M has been used in two ways: in defining a polygonal process, where we used the geodesics of the Levi-Civita-connection to get from q_t to q_{t+1}, i.e. for the edges of the polygon, and in defining the concept of independent directional variation, where the Riemannian metric was used to measure the length of the directional variation term.

3.6 Special Model of a polygonal process

Before we introduce the main model for a polygonal process of independent directional variation, we treat a special case of it. It will be the case that $E^{\mathcal{B}_{t-1}} d_t = 0$. This "special model" is the generalization of the usual random walk model for a coefficient process.

Special Model . Let M be a n-dimensional Riemannian manifold, TM its tangent bundle and let $m \in \mathbb{N}$, $m \le n$. For any $t \in \mathbf{Z}_+$ let be given m C^∞-vector fields $X_{t,1}, \cdots, X_{t,m}$ on M. Let $(\Omega, \mathcal{A}, P)$ be a probability space. Let $(\lambda_t)_{t \in \mathbf{Z}}$, $\lambda_t : \Omega \to \mathbf{R}^m$ be m-dimensional Gaussian white noise with variance $Q_t > 0$, and let $q_0 : \Omega \to M$ be $\mathcal{A}$-measurable, independent of the white noise $(\lambda_t)_{t \in \mathbf{Z}}$ and *non-evasive* i.e., such that

$$P(q_0 \in U) > 0 \text{ for all open sets } U \subset M. \tag{3.24}$$

We define the increasing sequence of σ-algebras $(\mathcal{B}_t)_{t \in \mathbf{Z}_+}$ by:

$$\mathcal{B}_t = \text{ the smallest } \sigma\text{-algebra, such that } q_0 \text{ and } \lambda_s (s \le t) \text{ are measurable,} \tag{3.25}$$

and we recursively define the process $(q_t, d_t)_{t\in\mathbf{Z}_+}, (q_t, d_t) : \Omega \to TM$ by:

$$d_t = \sum_{i=1}^{m} \lambda_{ti} X_{t,i} \tag{3.26}$$

$$q_{t+1} = \exp_{q_t}(d_t), \text{ if this is well defined. If not,} \tag{3.27}$$

$$q_{t+1} = h_{t+1} \text{ some } \mathcal{B}_t\text{-measurable } h_{t+1}. \tag{3.28}$$

Then, this is a polygonal process. Let ν be its running time. We shall show that the Never-Evasiveness Criterion and the Integrability Condition are satisfied, and that the directional variation term can normally be represented in vector fields.

The Never-Evasiveness Criterion: As the X_{ti} are continuous vector fields, the expressions q_s and d_s depend continuously on $q_0, \lambda_0, \cdots, \lambda_t$, if $s \leq t \leq \nu$. The proof of lemma 3.1, which shows that ν is a stopping time, now makes clear that the set in $M \times \mathbb{R}^{m(t+1)}$,

$$Y_t = \{(q_0, \lambda_0, \cdots, \lambda_t) \in M \times \mathbb{R}^{m(t+1)} |\ \nu > t-1 \text{ and } \sum_{s=0}^{t} ||d_s||_{q_s} < \varepsilon\}$$

is open, as well as its projection on $\mathbb{R}^{m(t+1)}$, the set Z_t,

$$Z_t = \{(\lambda_0, \cdots, \lambda_t) \in \mathbb{R}^{m(t+1)} |\ \text{there exists } q_0 \in M, \text{ such that } (q_0, \lambda_0, \cdot, \lambda_t) \in Y_t\}.$$

Then, for any open set $U \subset M$ we have:

$$P(q_0 \in U \text{ and } \sum_{s=0}^{t} ||d_s||_{q_s} < \varepsilon) \geq P(q_0 \in U \text{ and } (\lambda_0, \cdots, \lambda_t) \in Z_t) =$$

$$= P(q_0 \in U) P((\lambda_0, \cdots, \lambda_t) \in Z_t) > 0$$

because q_0 is non-evasive, $(\lambda_0, \cdots, \lambda_t)$ is Gaussian with non-singular variance, and Z_t is open.

The Integrability Condition: By the continuity of the vector fields $X_{t,i}$, the semi-positive definite matrix $(\langle X_{t,i}(q), X_{t,j}(q)\rangle_q)_{1\leq i,j\leq m}$ is bounded on any compact set $K \subset M$, and this implies that $\boldsymbol{E}(||d_t||_{q_t} 1_{\{q_t \in K\}}) < \infty$, because λ_t has finite second moments. Furthermore, we have $\boldsymbol{E}^{\mathcal{B}_{t-1}} d_t = 0$, because λ_t is independent of $\mathcal{B}_{t-1}$. Hence, the directional variation term is d_t, and can normally be represented in vector fields. Accordingly, we conclude that the process is of independent directional variation until ν if and only if the vector fields $X_{t,i}$ satisfy

$$q \mapsto \langle X_{t,i}(q), X_{t,j}(q)\rangle_q \text{ is constant on } M$$

for all $t \in \mathbf{Z}_+, 1 \leq i, j \leq m$. □

From now on we assume that the process is of independent directional variation until ν. In order to be useful later on in identification procedures, the transition equation

of the Model, (3.27) should be expressed in coordinates. We shall give an *approximate* expression: the expression of a (partly) *linearized* transition equation.
In the preceding corollary 3.7, we have seen that we can replace equation (3.26) by:

$$d_t = \sum_{k=1}^{k_t} \tilde{\lambda}_{t,k} \tau_{t,m_{t0}+k} F_{t,m_{t0}+k}, \tag{3.29}$$

where $0 < \tau_{t,m_{t0}+1} \leq \cdots \tau_{t,n}$, and $\tilde{\lambda}_t$ is unit variance Gaussian. Furthermore, $F_{t,m_{t0}+1}, \cdots, F_{t,n}$ is an everywhere orthonormal set of k_t C^∞ vector fields, and the $X_{t,i}$ can be expressed in the $F_{t,k}$ by means of coordinates, constant on the manifold. On any chart $(\tilde{U}_\alpha, x_\alpha)$, we can extend this orthonormal set to a base $(F_{t,k})_{1 \leq k \leq n}$, and then we can replace equation (3.26) by:

$$d_t = \sum_{j=1}^{n} \tilde{\lambda}_t \tau_{t,j} F_{t,j} \tag{3.30}$$

where $\tilde{\lambda}_t$ is n-dimensional unit variance Gaussian white noise and $0 \leq \tau_{t,1} \leq \cdots, \leq \tau_{t,n}$. The vector fields $F_{t,j}$ can be expressed in the basis $(\frac{\partial}{\partial x_{\alpha\ k}})_{1 \leq k \leq n}$:

$$F_{t,j}(p) = \sum_{k=1}^{n} g_{t,\alpha\ kj}(w) \frac{\partial}{\partial x_{\alpha\ k}} | p \text{ if } x_\alpha(p) = w.$$

Hence, the coordinates of $F_{t,j}$ with respect to this basis are in column number j of the invertible matrix $g_{t,\alpha}$. Let C_t be the n-dimensional diagonal matrix

$$C_t = \begin{pmatrix} \tau_{t,1} & 0 & \cdots & 0 \\ 0 & \tau_{t,2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \tau_{t,n} \end{pmatrix}.$$

Then,

$$d_t = \sum_{k=0}^{n} \sum_{i,j=1}^{n} g_{t,\alpha kj} C_{tji} \tilde{\lambda}_{ti} \frac{\partial}{\partial x_{\alpha\ k}}_{|q_t} = \sum_{k=0}^{n} (g_{t,\alpha} C_t \tilde{\lambda}_t)_k \frac{\partial}{\partial x_{\alpha\ k}}_{|q_t}. \tag{3.31}$$

Let H_α be the quadratic form, defining the Riemannian inner product on $\tilde{U}_\alpha$, i.e., let

$$(H_\alpha)_{ij}(w) = \langle \frac{\partial}{\partial x_{\alpha i}}, \frac{\partial}{\partial x_{\alpha j}} \rangle_p, \text{ if } x_\alpha(p) = w.$$

Because the $F_{t,j}$'s form an orthonormal base, we have:

$$g'_{t,\alpha} H_\alpha g_{t,\alpha} = I_n,$$

i.e.

$$g_{t,\alpha} g'_{t,\alpha} = H_\alpha^{-1}. \tag{3.32}$$

We express q_t in the coordinates of $\tilde{U}_\alpha$: let us denote

$$x_\alpha(q_t) =: v_t = \begin{pmatrix} v_{t1} \\ \vdots \\ v_{tn} \end{pmatrix}. \tag{3.33}$$

Now, we *linearize* the map $\exp_{q_t} : M_{q_t} \to M$ in the origin $0 \in M_{q_t}$, and then express the linearization of the transition equation

$$q_{t+1} = \exp_{q_t}(d_t) \tag{3.34}$$

in these coordinates. The derivative of the normal mapping in the origin is the identity (see chapter 2), and $\exp_{q_t}(0) = q_t$. The linearized transition equation (3.34), expressed in the coordinates x_α, gets the following form:

$$v_{t+1} = v_t + g_{t,\alpha}(v_t) C_t \tilde{\lambda}_t. \tag{3.35}$$

This linearized transition equation models the process $(v_t)_{t\in\mathbf{Z}_+}$, $v_t = x_\alpha(q_t)$ as a **martingale.**
Locally, there exists a chart, such that the linearized transition equation (3.35) expressed in the coordinates of this chart is *exactly* equivalent to the transition equation (3.34): use a *normal chart around* q_t. This is explained in chapter two. We shall discuss some conditions under which the system (3.35) is *globally* exactly equivalent with the transition equation of the process $(q_t, d_t)_{t\in\mathbf{Z}_+}$ in section (3.8)

3.7 Main model of a stochastic process on a Riemannian manifold.

The preceding Special Model is just a special case of our general model for a stochastic process on a Riemannian n dimensional manifold M of independent directional variation, according to its metric until the running time ν. We get this main model simply by replacing the unit variance white noise process $(\tilde{\lambda}_t)_{t\in\mathbf{Z}}$ in the preceding Special Model(3.30) by, more generally, a stationary Gaussian m-dimensional process $(\theta_t)_{t\in\mathbf{Z}}$.

General Model . Let M be a n-dimensional Riemannian manifold, TM its tangent bundle, and let $m \in \mathbb{N}$, $m \leq n$. For any $t \in \mathbf{Z}_+$ let be given m mutually orthonormal C^∞-vector fields $F_{t,n-m+1}, \cdots, F_{t,n}$ on M and an increasing sequence of m non-negative numbers $\tau_{t,n-m+1}, \cdots, \tau_{t,n}$. Let $(\Omega, \mathcal{A}, P)$ be a probability space. Let $(\theta_t)_{t\in\mathbf{Z}}$, $\theta_t : \Omega \to \mathbb{R}^m$, $\theta_t = (\theta_{t,n-m+1}, \cdots, \theta_{t,n})$ be an m-dimensional stationary process with unit conditional variance given the past, and a rational spectral density matrix $f(e^{i\phi})$ ($\phi \in [-\pi, \pi]$), which is everywhere invertible on $[-\pi, \pi]$. Let $q_0 : \Omega \to M$ be $\mathcal{A}$-measurable and independent of $(\theta_t)_{t\in\mathbf{Z}}$. We define the increasing sequence of σ-algebras $(\mathcal{B}_t)_{t\in\mathbf{Z}_+}$ by:

$$\mathcal{B}_t = \text{ the smallest } \sigma\text{-algebra, such that } q_0 \text{ and } \theta_s \ (s \leq t) \text{ are measurable.} \tag{3.36}$$

Then the process $(q_t, d_t)_{t\in\mathbf{Z}_+}$, $(q_t, d_t) : \Omega \to TM$ will be defined recursively by:

$$d_t = \sum_{j=n-m+1}^{n} \theta_{tj}\tau_{t,j}F_{t,j} \tag{3.37}$$

$$q_{t+1} = \exp_{q_t}(d_t), \text{if this is well defined. If not,} \tag{3.38}$$

$$q_{t+1} = h_{t+1}, \text{ some } \mathcal{B}_t\text{-measurable } h_{t+1}. \tag{3.39}$$

□

As the spectral density matrix f is rational, and everywhere non-singular, it is possible to factorize f in the form

$$f(e^{i\phi}) = \frac{1}{2\pi}k(e^{i\phi})Qk(e^{i\phi})^* \tag{3.40}$$

where $k(z)$ is rational in z, analytical within a circle containing the closed unit disk, $\det k(z) \neq 0$ for all $z, |z| < 1$, $k(0) = I$ and Q a positive definite n-dimensional matrix (Hannan and Deistler, Theorem 1.3.3). As the transfer function $k(z)$ is rational and analytical in a disk, containing the closed unit disk, we can make a stable state-space realization for the sequence $(\theta_t)_{t\in\mathbf{Z}}$:

$$\begin{aligned} \xi_{t+1} &= F\xi_t + K\lambda_t \\ \theta_t &= G\xi_t + \lambda_t \end{aligned} \tag{3.41}$$

(Hannan and Deistler, Theorem 1.2.1), where $(\lambda_t)_{t\in\mathbf{Z}}$ is white noise with variance Q. Since we have assumed that the spectral density matrix is invertible everywhere on $[-\pi, \pi]$, the transfer function $k(z) = G(I - Fz)^{-1}Kz + I$ is invertible on the closed unit disk, and it is a well known fact that this is possible if and only if besides F, also $F - KG$ is stable (just note that:

$$\begin{pmatrix} I - Fz & -Kz \\ G & I \end{pmatrix} =$$

$$= \begin{pmatrix} I & 0 \\ G(I - Fz)^{-1} & I \end{pmatrix} \begin{pmatrix} I - Fz & 0 \\ 0 & G(I - Fz)^{-1}Kz + I \end{pmatrix} \begin{pmatrix} I & -(I - Fz)^{-1}Kz \\ 0 & I \end{pmatrix}).$$

Consequently, instead of (3.41) we can also formulate the system as follows:

$$\begin{aligned} \xi_{t+1} &= (F - KG)\xi_t + K\theta_t \\ \lambda_t &= -G\xi_t + \theta_t, \end{aligned} \tag{3.42}$$

which again is a stable system. From this, it is easily seen that the smallest Hilbert space containing all θ_s, $s \leq t$, is equal to the smallest Hilbert space containing all λ_s, $s \leq t$, and, moreover, that $(\lambda_t)_{t\in\mathbf{Z}}$ is Gaussian. This implies that

$$d_t = \boldsymbol{E}^{\mathcal{B}_{t-1}}d_t + v_t, \text{ where}$$

$$\begin{aligned} \boldsymbol{E}^{\mathcal{B}_{t-1}}d_t &= \boldsymbol{E}^{\mathcal{B}_{t-1}}\Big(\sum_{i=n-m+1}^{n} \theta_{ti}\tau_{t,i}F_{t,i}(q_t)\Big) = \sum_{i=n-m+1}^{n} (G\xi_t)_i\tau_{t,i}F_{t,i}(q_t) \\ v_t &= \sum_{i=n-m+1}^{n} \lambda_{ti}\tau_{t,i}F_{t,i}(q_t), \end{aligned} \tag{3.43}$$

where λ_t is independent of $\mathcal{B}_{t-1}$. Moreover, we see that $Q = I$, because we have assumed that the conditional variance of θ_t given the past, is equal to the identity. From proposition 3.4 it is also immediately clear that the process is of independent directional variation. In order to write the model in coordinates, we can proceed as before: we extend the set of vector fields $F_{t,n-m+1}, \cdots, F_{t,n}$ to an orthonormal base $(F_{t,j})_{1 \leq j \leq n}$ on any chart $(\tilde{U}_\alpha, x_\alpha)$; we express these vector fields in coordinate matrices $g_{t,\alpha}$ w.r.t. the base $(\frac{\partial}{\partial x_{\alpha\, i}})_{1 \leq i \leq n}$, and linearize the transition equation $q_{t+1} = \exp_{q_t}(d_t)$. The variation terms v_t determine a symmetric tensor field V_t with eigenvalues $\tau_{t,i}$ and, possibly, zero, and furthermore the direction spaces $\mathcal{L}_t$. On the other hand, the stationary process θ, the direction spaces $\mathcal{L}_t$, and the symmetric tensor fields V_t, together with the initial probability distribution of q_0, determine the probability distribution of the whole process. In contrast with the special model, we now have to specify these direction spaces in general for this probability distribution, because the $\boldsymbol{E}^{\mathcal{B}_{t-1}}(d_t)$ is expressed in vector fields of $\mathcal{L}_t$. Let C_t be again the n-dimensional diagonal matrix with on its diagonal the increasing sequence of non-negative numbers $\tau_{t,i}$, and let θ_t be the n-dimensional stationary process $\theta_t = (0, \cdots, 0, \theta_{t,n-m+1}, \cdots, \theta_{t,n})$. Then the linearized transition equation $v_{t+1} = v_t + g_{t,\alpha}(v_t) C_t \theta_t$ in combination with (3.41) gives the system:

$$\begin{pmatrix} \xi_{t+1} \\ v_{t+1} \end{pmatrix} = \begin{pmatrix} F & 0 \\ g_{t,\alpha}(v_t) C_t G & I \end{pmatrix} \begin{pmatrix} \xi_t \\ v_t \end{pmatrix} + \begin{pmatrix} K \\ g_{t,\alpha}(v_t) C_t \end{pmatrix} \lambda_t. \qquad (3.44)$$

where

$$\begin{aligned} (\tilde{U}_\alpha, x_\alpha : \tilde{U}_\alpha &\to V_\alpha \subset \mathbf{R}^n) \text{ chart;} \\ x_\alpha(q_t) &= v_t \ ; \\ x_\alpha(q_{t+1}) &= v_{t+1} \text{ and} \\ E_{t,j}(q_t) &= \sum_{i=1}^{n} (g_{t,\alpha})_{ij}(v_t) \frac{\partial}{\partial x_{\alpha\, i}}. \end{aligned}$$

This system is almost linear, except for the non-linearity in the factor $g_{t,\alpha}(v_t)$. Furthermore, it is locally exactly equivalent to the transition equation of the process if one uses normal coordinates around the point q_t. In the next section (3.8), we will discuss the *global* equivalence between the transition equation and the linearized transition equation.

3.8 Equivalence of transition equation and its linearization

In this section, we shall investigate how linear our Model is, especially how linear its transition equation is. This is important, because, if the transition equation for the coefficients q_t of the time-varying AR(n) coefficient process $(y_t)_{t \in \mathbf{Z}_+}$ is linear in some coordinates, the model gets the character of a Dynamic Exponential Family Regression, an extension of Dynamic Generalized Linear Models. These are quite well studied in the statistical literature[4]. We shall examine the conditions that the transition equation is globally exactly equivalent with the transition equation of the process $(q_t, d_t)_{t \in \mathbf{Z}_+}$. These

[4]See e.g. McCullagh and Nelder (1989); Fahrmeir and Kaufmann(1991) and chapter six.

conditions depend on properties of the *distributions* $\mathcal{D}_t$, generated by the vector fields $F_{t,1}, \cdots, F_{t,m}$. In case we do not work with the Special Model, we also need a condition that the direction spaces $\mathcal{L}_t$ should satisfy.
We have already shown that dim $\mathcal{D}_t(p) = k_t$ for all $p \in M$ and some constant $k_t \in \mathbb{N}$. From proposition 3.6 it is clear that without loss of generality we may assume that $m = k_t$, which we shall do in this section. We need some well known results about *involutive* and *totally geodesic* distributions. See chapter two for these concepts.

Proposition 3.9 Suppose that the process $(q_t, d_t)_{t \in \mathbf{Z}_+}$, defined in the Model, is of independent directional variation until the running time ν, and that the distribution $\mathcal{D}_t$, generated by the vector fields in the direction space $\mathcal{L}_t$, is involutive for all $t \in \mathbf{Z}_+$. Then, the following statements are equivalent:

1. There exists a set of charts $(\tilde{U}_\alpha, x_\alpha : \tilde{U}_\alpha \to V_\alpha)$, covering M, and adapted to the distribution $\mathcal{D}_t$, such that for all α the transition equation of the process

$$q_{t+1} = \exp_{q_t}(d_t)$$

is exactly equivalent with (3.35) under the condition that $q_t, q_{t+1} \in \tilde{U}_\alpha$.

2. $\mathcal{D}_t$ is totally geodesic, and has curvature equal to zero.

Proof:

($1 \to 2$) Suppose that statement 1 is true. Let $F_{t,1}, \cdots, F_{t,m}$ be the moving orthonormal base in $\mathcal{L}_t$ as constructed before. Since the process is never-evasive, the transition equation $q_{t+1} = \exp_{q_t}(d_t)$, and $v_{t+1} = v_t + g_{t,\alpha} C_t \theta_t$ on the chart $(\tilde{U}_\alpha, x_\alpha)$, adapted to the distribution $\mathcal{D}_t$, are equivalent only if:

$$x_\alpha(\exp_q(\sum_{i=1}^{m} u_j F_{t,j})) = v + g_{t,\alpha}(v)u \text{ ,if } x_\alpha(q) = v \text{; for all } q \in \tilde{U}_\alpha,$$

$$\text{and all small } u = (u_1, \cdots, u_m, 0, \cdots, 0) \in \mathbf{R}^k. \tag{3.45}$$

Now, take a small $u = (u_1, \cdots, u_m, 0, \cdots, 0) \in \mathbf{R}^k$, $q \in \tilde{U}_\alpha$ and $t \in \mathbf{Z}_+$ arbitrarily. Let X be the vector field $X = \sum_{i=1}^{m} u_j F_{t,j}$; let $x_\alpha(q) = v \in V_\alpha$ and define the vector field $\tilde{X}$ by:

$$\tilde{X} = \sum_{i=1}^{m} (g_{t,\alpha}(v)u)_i \frac{\partial}{\partial x_{\alpha\, i}}. \tag{3.46}$$

Then, $X(q) = \tilde{X}(q)$. Moreover, because of (3.45), $\tilde{X}$ is the tangent field of a geodesic through q, hence we have:

$$(\nabla_{\tilde{X}} \tilde{X})\,(q) = 0.$$

As the coordinates $\tilde{X}_i = (g_{t,\alpha}(v)u)_i$ of $\tilde{X}$ with respect to the base $(\frac{\partial}{\partial x_{\alpha\, i}})_{1 \le i \le n}$ are constant on $\tilde{U}_\alpha$, we also have:

$$\nabla_{\tilde{X}} \tilde{X}\,(p) = \sum_{i,j=1}^{m} \tilde{X}_i(p) \tilde{X}_j(p) \nabla_{\frac{\partial}{\partial x_{\alpha\, i}}} \frac{\partial}{\partial x_{\alpha\, j}}\,(p) \text{ for all } p \in \tilde{U}_\alpha.$$

Using that $X(q) = \tilde{X}(q)$, we find:

$$\sum_{i,j=1}^{m} X_i(q)X_j(q)\nabla_{\frac{\partial}{\partial x_{\alpha\, i}}}\frac{\partial}{\partial x_{\alpha\, j}}(q) = 0$$

for all vectors $X(q) \in M_q$, such that $X(q) = \sum_{i=1}^{m} u_j F_{t,j}(q)$ for some $u = (u_1, \cdots, u_m, 0, \cdots, 0) \in \mathbb{R}^k$. But all vectors in $\mathcal{D}_t(q)$ can be written like that, and we conclude that

$$\nabla_{\frac{\partial}{\partial x_{\alpha\, i}}}\frac{\partial}{\partial x_{\alpha\, j}}(q) + \nabla_{\frac{\partial}{\partial x_{\alpha\, j}}}\frac{\partial}{\partial x_{\alpha\, i}}(q) = 0.$$

Because we also have:

$$\nabla_{\frac{\partial}{\partial x_{\alpha\, i}}}\frac{\partial}{\partial x_{\alpha\, j}}(q) = \nabla_{\frac{\partial}{\partial x_{\alpha\, j}}}\frac{\partial}{\partial x_{\alpha\, i}}(q),$$

it follows that

$$\nabla_{\frac{\partial}{\partial x_{\alpha\, i}}}\frac{\partial}{\partial x_{\alpha\, j}}(q) = 0 \text{ for all } q \in \tilde{U}_\alpha; 1 \leq i, j \leq m. \tag{3.47}$$

From this it is easily concluded that $\nabla_X Y \in \mathcal{D}_t$ holds for all vector fields $X, Y \in \mathcal{D}_t$. It also shows that the curvature tensor $R(X,Y)Z$ is zero for all vector fields X, Y and $Z \in \mathcal{D}_t$. Consequently, the distribution $\mathcal{D}_t$ is totally geodesic and has curvature zero.

$(2 \rightarrow 1)$ Choose an arbitrary point $q \in M$. Take a chart $(\tilde{U}_\alpha, x_\alpha : \tilde{U}_\alpha \rightarrow V_\alpha)$ in the neighbourhood of q, adapted to the distribution $\mathcal{D}_t$. Adaptedness means that $\mathcal{D}_t$ is spanned by some coordinate base vectors, say by $\frac{\partial}{\partial x_{\alpha\, 1}}, \cdots, \frac{\partial}{\partial x_{\alpha\, m}}$. We shall construct m independent vector fields $B_1, \cdots, B_m$ in $\mathcal{D}_t$ around q, such that $\nabla_X B_j = 0$ for all $j, 1 \leq j \leq m$, and all vector fields $X \in \mathcal{D}_t$.
We simply show that the partial differential equation

$$\nabla_{\frac{\partial}{\partial x_{\alpha\, j}}} Z = 0 \text{ for all } 1 \leq j \leq m \tag{3.48}$$

is solvable in $\mathcal{D}_t$ for every starting condition $Z(q) = w \in \mathcal{D}_t(q)$. To see this, we first express Z w.r.t. an orthonormal base $(F_{t,j})_{1 \leq j \leq m}$ of $\mathcal{L}_t$:

$$Z = \sum_{l=1}^{m} Z_l F_{t,l}.$$

Using this, equation (3.48) can be rewritten as:

$$\sum_{l=1}^{m} \frac{\partial}{\partial x_{\alpha\, j}}(Z_l)F_{t,l} + \sum_{l=1}^{m} Z_l \nabla_{\frac{\partial}{\partial x_{\alpha\, j}}} F_{t,l} = 0.$$

By taking the inner product with $F_{t,k}$, we get

$$\frac{\partial}{\partial x_{\alpha\, j}}(Z_k) = -\sum_{l=1}^{m} Z_l \langle F_{t,k}, \nabla_{\frac{\partial}{\partial x_{\alpha\, j}}} F_{t,l}\rangle \text{ for all } 1 \leq k, j \leq m,$$

or, in vector notation:

$$\frac{\partial}{\partial x_{\alpha\, j}}(Z) = -A_j Z \text{ for all } 1 \leq j \leq m, \tag{3.49}$$

where

$$A_j = \begin{pmatrix} \langle F_{t,1}, \nabla_{\frac{\partial}{\partial x_{\alpha\, j}}} F_{t,1}\rangle & \cdots & \langle F_{t,1}, \nabla_{\frac{\partial}{\partial x_{\alpha\, j}}} F_{t,m}\rangle \\ \vdots & \vdots & \vdots \\ \langle F_{t,m}, \nabla_{\frac{\partial}{\partial x_{\alpha\, j}}} F_{t,1}\rangle & \cdots & \langle F_{t,m}, \nabla_{\frac{\partial}{\partial x_{\alpha\, j}}} F_{t,m}\rangle \end{pmatrix}.$$

Such a linear partial differential equation with starting condition $Z(q) = w \in \mathbb{R}^m$ has a solution, *provided* differentiating equation (3.49) entry j with respect to $x_{\alpha\, i}$ will give the same equation as differentiating equation (3.49) entry i with respect to $x_{\alpha\, j}$. This condition, which correctness has to be shown, can be reformulated as

$$-\frac{\partial}{\partial x_{\alpha i}} A_j + A_j A_i = -\frac{\partial}{\partial x_{\alpha j}} A_i + A_i A_j \tag{3.50}$$

We show this equality by evaluating entry kr of this equation. We have:

(a)

$$\left(\frac{\partial}{\partial x_{\alpha\, i}} A_j\right)_{kr} = \frac{\partial}{\partial x_{\alpha\, i}} \langle F_{t,k}, \nabla_{\frac{\partial}{\partial x_{\alpha\, j}}} F_{t,r}\rangle =$$

$$= \langle \nabla_{\frac{\partial}{\partial x_{\alpha i}}} F_{t,k}, \nabla_{\frac{\partial}{\partial x_{\alpha j}}} F_{t,r}\rangle + \langle F_{t,k}, \nabla_{\frac{\partial}{\partial x_{\alpha i}}} \nabla_{\frac{\partial}{\partial x_{\alpha\, j}}} F_{t,r}\rangle, \text{ and}$$

(b) because

$$0 = \frac{\partial}{\partial x_{\alpha j}} \langle F_{t,k}, F_{t,l}\rangle = \langle \nabla_{\frac{\partial}{\partial x_{\alpha j}}} F_{t,k}, F_{t,l}\rangle + \langle F_{t,k}, \nabla_{\frac{\partial}{\partial x_{\alpha j}}} F_{t,l}\rangle,$$

and

$$\mathcal{D}_t \ni \nabla_{\frac{\partial}{\partial x_{\alpha\, j}}} F_{t,k} = \sum_{l=1}^{m} \langle \nabla_{\frac{\partial}{\partial x_{\alpha\, j}}} F_{t,k}, F_{t,l}\rangle F_{t,l},$$

we also have:

$$(A_j A_i)_{kr} = \sum_{l=1}^{m} (A_j)_{kl} (A_i)_{lr} = \sum_{l=1}^{m} \langle F_{t,k}, \nabla_{\frac{\partial}{\partial x_{\alpha\, j}}} F_{t,l}\rangle \langle F_{t,l}, \nabla_{\frac{\partial}{\partial x_{\alpha\, i}}} F_{t,r}\rangle =$$

$$= -\sum_{l=1}^{m} \langle \nabla_{\frac{\partial}{\partial x_{\alpha\, j}}} F_{t,k}, F_{t,l}\rangle \langle F_{t,l}, \nabla_{\frac{\partial}{\partial x_{\alpha\, i}}} F_{t,r}\rangle = -\langle \nabla_{\frac{\partial}{\partial x_{\alpha j}}} F_{t,k}, \nabla_{\frac{\partial}{\partial x_{\alpha\, i}}} F_{t,r}\rangle.$$

Since the curvature is zero in $\mathcal{D}_t$, it follows that

$$\nabla_{\frac{\partial}{\partial x_{\alpha\, j}}} \nabla_{\frac{\partial}{\partial x_{\alpha\, i}}} F_{t,r} = \nabla_{\frac{\partial}{\partial x_{\alpha\, i}}} \nabla_{\frac{\partial}{\partial x_{\alpha\, j}}} F_{t,r},$$

which shows the correctness of condition (3.50).
Hence, we have m vector fields $B_1, \cdots B_m$ in $\mathcal{D}_t$, independent in M_q, and thus also independent in the neighbourhood of q, such that $\nabla_X B_i = 0$ for all $X \in \mathcal{D}_t$. Then, in particular we have $[B_i, B_j] = \nabla_{B_i} B_j - \nabla_{B_j} B_i = 0$ and, according to Nijmeijer

and van der Schaft Theorem 2.36, we can find a chart $(\tilde{U}_\beta, x_\beta : \tilde{U}_\beta \to V_\beta \subset \mathbb{R}^n)$, such that $B_i = \frac{\partial}{\partial x_{\beta i}}$ for all $1 \leq i \leq m$. Let again $(F_{t,j})_{1 \leq j \leq m}$ be an orthonormal base of eigenvectors of the symmetric tensor V_t in the direction space $\mathcal{L}_t$. Then, the transition equation of the Model at time t is

$$q_{t+1} = \exp_{q_t}(\sum_{i=1}^{m} \lambda_{t,i}\tau_{t,i}F_{t,i}). \tag{3.51}$$

Let $x_\beta(q_t) = (v_{t1}, \cdots, v_{tk}) = v_t$. The vector field $F_{t,i}$ can be expressed in the coordinate vectorfields $\frac{\partial}{\partial x_{\beta\ j}} = B_j$, $(1 \leq j \leq m)$ by means of the coefficients in the ith column of matrix $g_{t,\beta}$. Hence, the vector $\tilde{X}(q_t) = \sum_{i=1}^{m} \lambda_{t,i}\tau_{t,i}F_{t,i}$ in the point q_t can be written as:

$$\tilde{X}(q_t) = \sum_{j=1}^{m} (g_{t,\beta}(v_t)u)_j B_j$$

where $u = (\lambda_{t,1}\tau_{t,1}, \cdots, \lambda_{t,m}\tau_{t,m}, 0, \cdots, 0) \in \mathbb{R}^k$. As we have $\nabla_{B_i} B_j = 0$, the vector field $\tilde{X}$, defined by $\tilde{X} = \sum_{j=1}^{m} (g_{t,\beta}(v_t)u)_j B_j$ is a tangent field of a geodesic. Consequently, the translation of equation (3.51) on chart $(\tilde{U}_\beta, x_\beta)$ is *exactly*

$$(v_{t+1\ 1}, \cdots, v_{t+1\ k}) = (v_{t1}, \cdots, v_{tk}) + g_{t,\beta}(v_t)u,$$

where $x_\beta q_{t+1} = v_{t+1}$ and u is as defined above. □

Example 3.10 A simple example of an involutive, totally geodesic distribution with curvature zero is a distribution generated by the *field of tangent vectors of all the geodesics through a certain point.*
The precise construction is as follows. Start with an arbitrary Riemannian manifold $\tilde{M}$, choose a point $q_{-1} \in \tilde{M}$, and likewise an open *normal* neighbourhood U (with normal property, see chapter two) around the point q_{-1}. Now, let M be the Riemannian manifold:

$$M = U \setminus q_{-1}.$$

Any point $p \in M$ can be connected with q_{-1} by one and only one geodesic γ in U, such that

$$\gamma(0) = q_{-1}\ ;\ \dot{\gamma}(0) = v \text{ with } ||v||_{q_{-1}} = 1 \text{ and } \gamma(t) = p \text{ for some } t \in \mathbb{R}.$$

This implies that

$$\exp_{q_{-1}}(tv) = p.$$

Now, define the vector field F by:

$$F(p) = \dot{\gamma}(t).$$

Then, F is the tangent field of the geodesics through the point q_{-1}. It can be verified that F is a C^∞-vector field with constant length one. The distribution generated by F

is constant dimensional, involutive, and has curvature zero (all one-dimensional distributions have curvature zero and are involutive), and totally geodesic.
We shall often make use of this example for constructing a very interesting type of Models: the Geodesic Models.
Definition. A General Model on a manifold M will be called a *Geodesic Model* if, for some geodesic tangent field F on M, all direction spaces $\mathcal{L}_t$ satisfy

$$\mathcal{L}_t = \boldsymbol{span}\{F\}. \tag{3.52}$$

Instead of a tangent field F of geodesics through a point, one can also make tangent fields of geodesics perpendicular to a hypersurface. In chapter five we shall encounter two examples of fields of the latter type:

- Let $M = U_n$, the stability region for associated polynomials of stationary AR(n) processes, and let the Riemann structure on M be given by the asymptotic Fisher metric. On this manifold we have the example of the vector field $F = \frac{\partial}{\partial \alpha_n}$, the geodesic tangent field given by the arcsines of the last Schur parameter. It is perpendicular to the hypersurface U_{n-1}.

- On the same Riemannian manifold, we also have the example of $F = \frac{\partial}{\partial r}$, the geodesic tangent field given by the log-noise-level. It is perpendicular to the hypersurface AR_n, of polynomials corresponding to AR(n) processes with noise-level equal to one. □

Under which circumstances can we obtain a linearized transition equation that is not only exact but also completely linear? The answer is given in the following Corollary. We call it "Corollary", because it is proved in almost the same way as the preceding proposition.

Corollary 3.11 Suppose that the process $(q_t, d_t)_{t\in\mathbb{Z}_+}$, defined in the Model on the k-dimensional manifold M, is of independent directional variation until the running time ν with direction spaces $\mathcal{L}_t$ that generate distributions $\mathcal{D}_t (t \in \mathbb{Z}_+)$.
Then, there exists a set of charts, covering M, and adapted to the distribution $\mathcal{D}_t$, such that the transition equation at time t, expressed in coordinates of these charts, is not only exact but also completely linear *if and only if* all vector fields X and Y in the direction space $\mathcal{L}_t$ satisfy $\nabla_X Y = 0$.

Proof: Let $F_{t,1}, \cdots, F_{t,m}$ be an orthonormal base of the direction space $\mathcal{L}_t$. As the exact transition equation is equivalent to its linearized one, we have:

$$x_\alpha(\exp_q(\sum_{i=1}^{m} u_j F_{t,j})) = v + g_{t,\alpha}(v)u, \text{ if } x_\alpha(q) = v$$

for all $q \in \tilde{U}_\alpha$ and all small $u = (u_1, \cdots, u_m, 0, \cdots 0) \in \mathbb{R}^k$ where, as we have seen in the proof of the previous proposition 3.9, the first m coordinate base vector fields $\frac{\partial}{\partial x_{\alpha\, i}}$ $(1 \leq i \leq m)$ satisfy

$$\nabla_{\frac{\partial}{\partial x_{\alpha\, i}}} \frac{\partial}{\partial x_{\alpha\, j}} = 0 \quad \text{(see (3.47))}.$$

The transition equation is completely linear, hence the first m columns of the matrices $g_{t,\alpha}(v)$ are constant on V_α; these do not depend on v. Consequently, the m vector fields $F_{t,j}$ are constant linear combinations of the first m coordinate base vector fields. The elements X, Y of the direction space $\mathcal{L}_t$ on their turn are constant linear combinations of the $F_{t,j}$, hence also of the first m coordinate base vector fields, and, accordingly, X and Y satisfy $\nabla_X Y = 0$.
Suppose on the other hand that all elements X and Y of the direction space $\mathcal{L}_t$ satisfy $\nabla_X Y = 0$. The rest of the proof is almost a repetition of the last paragraph of the proof of the preceding proposition: Let $F_{t,1}, \cdots, F_{t,m}$ be an orthonormal base of eigenvectors in $\mathcal{L}_t$ of the symmetric tensor V_t. In particular we have $[F_{t,i}, F_{t,j}] = \nabla_{F_{t,i}} F_{t,j} - \nabla_{F_{t,j}} F_{t,i} = 0$. We can apply proposition 2.36 from Nijmeier and van der Schaft: around any point $q \in M$, there exists a chart $(\tilde{U}_\beta, x_\beta)$, such that

$$\frac{\partial}{\partial x_{\beta\, i}} = F_{t,i} \qquad (1 \leq i \leq m),$$

hence, adapted to the distribution $\mathcal{D}_t$. Any constant linear combination X of the $F_{t,i}$ is in $\mathcal{L}_t$, hence satisfies $\nabla_X X = 0$, and, accordingly, is a tangent field of a geodesic. The Model defines the following transition equation:

$$q_{t+1} = \exp_{q_t}(\sum_{i=1}^{m} \lambda_{t,i} \tau_{t,i} F_{t,i}).$$

As $X = \sum_{i=1}^m \lambda_{t,i}\tau_{t,i}F_{t,i}$ is the tangent field of a geodesic on the chart $(\tilde{U}_\beta, x_\beta)$, we exactly get:

$$(v_{t+1\ 1}, \cdots, v_{t+1\ k}) = (v_{t1}, \cdots, v_{tk}) + (\lambda_{t,1}\tau_{t,1}, \cdots, \lambda_{t,m}\tau_{t,m}, 0, \cdots, 0)$$

with

$$x_\beta(q_{t+1}) = (v_{t+1\ 1}, \cdots, v_{t+1\ k}); \quad x_\beta(q_t) = (v_{t1}, \cdots, v_{tk}).$$

The right hand side of this equation is linear in λ_t and v_t. □

Note that the probability distribution of a Special Model only depends on the eigen *distributions* of the symmetric tensors V_t. If all the eigen distributions corresponding to non-zero eigenvalues $\tau_{t\ i}$ of V_t are totally geodesic, and $\mathcal{D}_t$ has curvature zero, then the Special Model is observationally equivalent to a Model with transition equation at time t that on some chart is exactly equivalent to a completely linear equation.

As we already said in the first paragraph of this section, the study of models with a linear transition equation will take an important place in this thesis. The reader will not be surprised about the word we use for this type of models.

Definition. A zero-curvature model is a General Model on a manifold M, such that M can be covered by charts with respect to which the transition equation is exactly linear. □

A fairly rough estimate of the error, made in replacing the transition equation of the process at time t by (3.35), is contained in the following lemma 3.12. First, we introduce

some notation.

Notation
Let $(\tilde{U}_\alpha, x_\alpha : \tilde{U}_\alpha \to V_\alpha \subset \mathbb{R}^n)$ be a chart of a n-dimensional Riemannian manifold M. In every point $w \in V_\alpha$, we define the following $n+1$ symmetric n-dimensional square matrices: if $x_\alpha(p) = w$ and $1 \le i, j, k \le n$ then

$$\begin{aligned} \text{the matrix defining the } \textit{metric is: } (H_\alpha)_{ij}(w) &:= \langle \frac{\partial}{\partial x_{\alpha\, i}}, \frac{\partial}{\partial x_{\alpha\, j}} \rangle_{|p}; \\ \text{the } \textit{christoffel symbols } (\Gamma^k)_{ij}(w) \text{ are} &: \text{matrices defined by} \\ \nabla_{\frac{\partial}{\partial x_{\alpha\, i}}} \frac{\partial}{\partial x_{\alpha\, j}}(p) &= \sum_{k=1}^{n} \Gamma^k_{i,j}(w) \frac{\partial}{\partial x_{\alpha\, k}}(p). \end{aligned} \quad (3.53)$$

These matrices are smooth functions of w. Furthermore, for any compact set $K \subset V_\alpha$, we define the non-negative numbers

$$\begin{aligned} \Gamma_{\max}(K) &= \max\{|\lambda| \mid \lambda \text{ eigenvalue of } \Gamma^k(w),\ 1 \le k \le n,\ w \in K\} \\ H_{\max}(K) &= \max\{\mu \mid \mu \text{ eigenvalue of } H_\alpha(w),\ w \in K\} \\ H_{\min}(K) &= \min\{\mu \mid \mu \text{ eigenvalue of } H_\alpha(w),\ w \in K\}. \end{aligned} \quad (3.54)$$

Lemma 3.12 Let $(E_j)_{1 \le j \le n}$ be a moving orthonormal base of the tangent spaces of M,

$$E_j(p) = \sum_{i=1}^{n} (g_\alpha)_{ij}(w) \frac{\partial}{\partial x_{\alpha\, i}}, \text{ if } x_\alpha(p) = w,$$

and let q be a point in $\tilde{U}_\alpha$; $x_\alpha(q) = v$. Then, there is a compact neighbourhood W of v in V_α and a number $r > 0$, such that

$$d(\exp_q(\sum_{i=1}^{n} u_j E_j), x_\alpha^{-1}(v + g_\alpha(v)u)) \le \frac{\sqrt{n}\sqrt{H_{\max}(W)}\Gamma_{\max}(W)}{2H_{\min}(W)} \sum_{j=1}^{n} u_j^2 \quad (3.55)$$
$$\text{for all } u \in \mathbb{R}^n \text{ satisfying } \sum_{j=1}^{n} u_j^2 < r^2.$$

Proof: We use Theorem I.9.9, Helgason (see chapter two) to make a compact ball $B_{r_q}(q)$ around q, such that its image $x_\alpha(B_{r_q}(q))$ lies in a ball with radius $\tilde{r}$ in V_α, and such that any two points in $B_{r_q}(q)$ can be connected by a geodesic in $B_{r_q}(q)$ with length equal to the distance of these two points. Let W be the convex closure of $x_\alpha(B_{r_q}(q))$ and let $r = \min(r_q, \tilde{r})$. Then, W is compact.
Firstly, we approximate the distance of two points $p, \tilde{p}$ in $B_{r_q}(q)$ by the Euclidean distance of their corresponding points $w, \tilde{w}$ in W. Connect the points $w, \tilde{w}$ by a straight line in W; let $\eta : [0,1] \to M; \eta(0) = p; \eta(1) = \tilde{p}$ be the corresponding curve on M. Then,

$$d(p, \tilde{p}) \le \int_0^1 ||\dot{\eta}(t)||_{\eta(t)} dt \le \sqrt{H_{\max}(W)} ||w - \tilde{w}||_{\text{Eucl}}. \quad (3.56)$$

Secondly, we compare, first in V_α, the image $\gamma = x_\alpha(\tilde{\gamma})$ of the geodesic $\tilde{\gamma} : [0,1] \to M; \tilde{\gamma}(t) = \exp_q(t \sum_{j=1}^{n} u_j E_j)$ with the straight line

$\delta : [0,1] \to V_\alpha; \delta(t) = v + g_\alpha(v)ut$. Let $h : [0,1] \to V_\alpha; h(t) = \gamma(t) - \delta(t)$ be the difference of these two curves in V_α. Then,

$$h(0) = 0; h'(0) = 0; h''(t) = \ddot{\gamma}(t) \text{ for all } t \in [0,1],$$

and we have:

$$||h(t)||_\infty = ||\int_0^t \int_0^s h''(\tau)d\tau dt||_\infty = ||\int_0^t (t-\tau)h''(\tau)d\tau||_\infty \leq \frac{t^2}{2} \max_{\tau\in[0,t]} ||h''(\tau)||_\infty.$$

Because $\tilde{\gamma}$ is a geodesic, we have:

$$\ddot{\gamma}_k(\tau) = -\sum_{i,j=1}^n \dot{\gamma}_i(\tau)\Gamma_{ij}^k(\tau)\dot{\gamma}_j(\tau),$$

and, consequently,

$$|\ddot{\gamma}_k(\tau)| \leq \Gamma_{\max}(W)||\dot{\gamma}(\tau)||^2_{\text{Euc}}.$$

Now, we compare the Euclidean norm $||\ ||_{\text{Euc}}$ on V_α with the Riemannian norm $||\ ||_{\tilde{\gamma}(\tau)}$:

$$||\dot{\gamma}(\tau)||^2_{\text{Euc}} = \langle \dot{\gamma}(\tau), H_\alpha(\gamma(\tau))H_\alpha^{-1}(\gamma(\tau))\dot{\gamma}(\tau)\rangle \leq \frac{1}{H_{\min}(W)}||\dot{\tilde{\gamma}}(\tau)||^2_{\tilde{\gamma}(\tau)}.$$

The tangent vector of a geodesic has constant length, hence we have:

$$||\dot{\tilde{\gamma}}(\tau)||^2_{\tilde{\gamma}(\tau)} = \sum_{j=1}^n u_j^2.$$

It follows that

$$||\ddot{\gamma}(\tau)||_\infty \leq \frac{\Gamma_{\max}(W)}{H_{\min}(W)} \sum_{j=1}^n u_j^2.$$

Inequality (3.55) now follows immediately from

$$||x||_{\text{Euc}} \leq \sqrt{n}||x||_\infty \text{ for all } x \in \mathbb{R}^n.$$

□

Remark 3.13 It is a well known fact (Kobayashi and Nomizu,I,1963, Proposition 8.4) that for a *normal* chart (U_α, x_α) around a point p, as defined in chapter two, the christoffel symbols vanish in the point p, i.e.

$$\Gamma_{ij}^k(w) = 0 \text{ if } x_\alpha(p) = w.$$

With this fact, the inequalities of the lemma show that normal charts have locally good properties as far as the equivalence of transition equation and its linearized version is concerned. □

3.9 Complexity

The design of a Special Model of independent directional variation until the running time ν depends on the following specifications:

1. The specification of the Riemannian manifold M;

2. The specification of the probability distribution of the initial point $q_0 : \Omega \to M$;

3. The specification of the vector fields $X_{t,1}, \cdots, X_{t,m}$ and the variance of λ_t, $Q_t > 0$ for all $t \in \mathbf{Z}_+$.

As we have seen, we can replace point 3. by:

3.' The specification for all $t \in \mathbf{Z}_+$, of a finite, increasing sequence of positive numbers $\tau_{t,j}$, and
corresponding with these, a finite sequence of mutually orthonormal C^∞ vector fields $F_{t,j}$.

On any chart $(\tilde{U}_\alpha, x_\alpha)$, we can extend the set of vector fields $F_{t,j}$ to an orthonormal base $(F_{t,j})_{1 \le j \le n}$, such that the last k_t vector fields $F_{t,n-k_t+1}, \cdots, F_{t,n}$ span the same distribution $\mathcal{D}_t$ as the vector fields $X_{t,i}$. Expressing the moving orthonormal base $(F_{t,j})_{1 \le j \le n}$ on the coordinate neighbourhood $\tilde{U}_\alpha$ in the coordinates yields a factorization of the inverse of the quadratic form H_α, as equation (3.32) shows. At the intersection of two coordinate neighbourhoods $\tilde{U}_\alpha \cap \tilde{U}_\beta$, we have:

$$F_{t,j} = \sum_{r=1}^{n} g_{t,\beta r j} \frac{\partial}{\partial x_{\beta,r}} = \sum_{r,k=1}^{n} g_{t,\beta r j} (D x_\alpha \circ x_\beta^{-1})_{kr} \frac{\partial}{\partial x_{\alpha,k}} \quad \text{for all } 1 \le j \le k_t,$$

hence,

there is a number $k_t \in \mathbf{Z}_+$, such that for all charts $(\tilde{U}_\alpha, x_\alpha)$, $(\tilde{U}_\beta, x_\beta)$ and for all $p \in \tilde{U}_\alpha \cap \tilde{U}_\beta$, the last k_t columns of $g_{t,\alpha}(w)$ are equal to the last k_t columns of $(D x_\alpha \circ x_\beta^{-1}) g_{t,\beta}(u)$ if $x_\alpha(p) = w$ and $x_\beta(p) = u$.

(3.57)

It is not difficult to verify that, if a collection of factorizations $H_\alpha^{-1} = g_{t,\alpha} g'_{t,\alpha}$ on all coordinate neighbourhoods $\tilde{U}_\alpha$ satisfies the compatibility condition (3.57), then this collection completely determines a moving orthonormal set $F_{t,n-k_t+1}, \cdots, F_{t,n}$ on M, and a moving orthonormal base on any chart. Hence, we can replace point 3'. by:

3". For any $t \in \mathbf{Z}_+$, one has to specify an increasing sequence of n non-negative numbers $\tau_{t,1}, \cdots, \tau_{t,n}$ and a collection of charts $(\tilde{U}_\alpha, x_\alpha)$, covering M, together with factorizations $H_\alpha^{-1} = g_{t,\alpha} g'_{t,\alpha}$ satisfying the compatibility condition (3.57).

If we are only interested in the probability distribution of the process, then we merely have to specify:

1. The Riemannian manifold M,

2. The probability distribution of the initial point $q_0 : \Omega \to M$,

3. For any $t \in \mathbf{Z}_+$, the symmetric tensor V_t of type $(2,0)$ with non-negative eigenvalues and multiplicities which are constant on the manifold.

As we stated in the remark 3.5, we can specify instead of point number 3:

3"'. For any $t \in \mathbf{Z}_+$, a finite set $\mathcal{T}_t$ of non-negative numbers τ_{ti} and, corresponding with these, a set of constant dimensional distributions $\mathcal{D}_{t\,\tau_{ti}}$, which are mutually orthogonal, and which together span $\mathcal{D}M$.

Complexity of a model depends on the number of specifications to be made. In order to reduce the complexity of a Model, one can make a priori assumptions on the process like e.g. that the vector fields $X_{t,i}$, or the tensors V_t, are all equal to 0 except in one "jump" time, or that the vector fields do not depend on the time t. In the latter case that the vector fields are constant in time, we can now give a list of *Special* Models of increasing complexity:

Complexity types

1. The set $\mathcal{T}$ of non-negative numbers mentioned in point 3"'. contains only one number τ that, consequently, has to correspond to the total distribution $\mathcal{D}M$. The probability distribution (until ν) depends only on that number τ and, as we shall see in the applications, in a not very critical way on the probability distribution of q_0.
   We shall call these models the *Simple Models.*

2. The set $\mathcal{T}$ contains only two numbers, $\tau_0 = 0$ and $\tau > 0$, where τ corresponds to an involutive, totally geodesic distribution $\mathcal{D}_\tau$ with constant dimension m. Then, the probability distribution depends on the number τ^2, the distribution $\mathcal{D}_\tau$, and also on the probability distribution of q_0 in a critical way. In fact, q_0 determines on which integral manifold $N(q_0)$ of $\mathcal{D}_\tau$ the process will proceed until the time ν. According to Frobenius' Theorem (see Nijmeijer and Van der Schaft, Corollary 2.43), for any integral manifold N of an involutive distribution and any chart $(\tilde{U}_\alpha, x_\alpha)$, adapted to the distribution, there are numbers $a_1, \cdots, a_{n-m} \in \mathbf{R}$, such that

$$N \cap \tilde{U}_\alpha = \{q \in \tilde{U}_\alpha | x_{\alpha\, m+1}(q) = a_1, \cdots, x_{\alpha\, n}(q) = a_{n-m}\}.$$

   Determining the manifold $N(q_0)$ is the same as determining $x_{\alpha\, m+1}(q_0), \cdots, x_{\alpha\, n}(q_0)$ if $q_0 \in \tilde{U}_\alpha$.

3. The set $\mathcal{T}$ consists of two positive numbers τ_1 and τ_2. The distribution $\mathcal{D}_{\tau_1}$ is constant dimensional and involutive, and the distribution $\mathcal{D}_{\tau_2}$ is of constant dimension 1 (so it is involutive), and totally geodesic. Although this is already a very complicated situation[5], we can construct charts that reduce the complexity. It is done in the following lemma 3.14. The lemma is almost trivial if one is familiar with *pencils of geodesics*. The definition of pencils of geodesics, and some well known facts about these, can be found in chapter two.

Lemma 3.14 Consider a Model with tensors V_t ($t \in \mathbf{Z}_+$). Suppose that for some time $t \in \mathbf{Z}_+$ the tensor V_t is characterized by only two eigenvalues $0 \leq \tau_{t1} < \tau_{t2}$ with multiplicities $n-1$ and 1 respectively, and corresponding orthogonal constant dimensional eigen distributions $\mathcal{D}_{\tau_{t1}}$ and $\mathcal{D}_{\tau_{t2}}$. Then, the following statements are equivalent.

1. There is a collection of charts $(U_\eta, x_\eta : U_\eta \to V_\eta)$ covering M, such that

   (a) They are adapted to both distributions $\mathcal{D}_{\tau_{t1}}$ and $\mathcal{D}_{\tau_{t2}}$,

   (b) The linearization of the transition equation $q_{t+1} = \exp_{q_t}(d_t)$ on this chart is:

$$\begin{aligned}
\text{if } x_\eta(q_t) &= \begin{pmatrix} v_{t\,1} & \cdots & v_{t\,n-1} & w_t \end{pmatrix}' \text{ and,} \\
\text{if } x_\eta(q_{t+1}) &= \begin{pmatrix} v_{t+1\,1} & \cdots & v_{t+1\,n-1} & w_{t+1} \end{pmatrix}' \text{ then,} \\
v_{t+1} &= v_t + \tilde{g}_\eta(v_t, w_t)\tau_{t1}\tilde{\lambda}_t, \qquad (3.58) \\
w_{t+1} &= w_t + \tau_{t2}\lambda_t, \qquad (3.59)
\end{aligned}$$

   where $\begin{pmatrix} \tilde{\lambda}_t \\ \lambda_t \end{pmatrix}$ is unit variance white noise, and
   $g(v,w) = \begin{pmatrix} \tilde{g}_\eta(v,w) & 0_{n-1\times 1} \\ 0_{1\times n-1} & 1 \end{pmatrix}$ satisfies $gg' = H_\eta^{-1}$
   ($H_\eta(v,w)$ is the matrix of the Riemannian inner product on the coordinate bases of the tangent spaces of U_η).

   (c) Equation (3.59) is precise.

2. The distribution $\mathcal{D}_{\tau_{t1}}$ is involutive, and the distribution $\mathcal{D}_{\tau_{t2}}$ is totally geodesic.

Proof :

($1 \to 2$) One can deduce V_t from equations (3.58, 3.59); it is clear that the eigen distribution $\mathcal{D}_{\tau_{t1}}$ is spanned by the (orthonormal) vector fields $F_{t,1}, \cdots, F_{t,n-1}$. They have their coordinates w.r.t. base $(\frac{\partial}{\partial x_{\eta\,i}})_{1\leq i\leq n}$ in the first $n-1$ columns of the matrix g. As the last row of these columns is zero, the fields $F_{t,i}$ ($1 \leq i \leq n-1$) are already linear combinations of the $\frac{\partial}{\partial x_{\eta\,j}}$, ($1 \leq j \leq n-1$). Thus, $\mathcal{D}_{t1}$ is spanned

[5] One might argue that this case is simpler than case 2, if $\tau_1 > 0$, because then the model does not depend that heavily on the probability distribution of q_0.

by these first $n-1$ coordinate base vectors, and is hence involutive (A distribution spanned by coordinate base vector fields is always involutive). The other distribution $\mathcal{D}_{t2}$ is automatically involutive, because it is one-dimensional. Moreover, because equation (3.59) is precise, proposition 3.9 shows that $\mathcal{D}_{t2}$ is totally geodesic.

$(2 \to 1)$ We shall construct a chart with properties (a),(b),(c) around any point in M. The distribution $\mathcal{D}_{\tau_2}$ is spanned by a geodesic vector field B, i.e. $\nabla_B B = 0$. Then,

$$B\langle B,B\rangle_q = 2\langle \nabla_B B, B\rangle_q = 0 \text{ for all } q \in M, \text{ so also the vector field}$$

$$X(q) = \frac{B(q)}{||B||_q} \text{ satisfies } \nabla_X X = 0,$$

and is a geodesic vector field spanning the distribution $\mathcal{D}_{\tau_2}$.
Fix a point $p \in M$. Then, there exists a ball $B_{r_p}(p)$ with center p and radius r_p, such that any two points in it can be joined by a geodesic that lies in $B_{r_p}(p)$ (normal property, see chapter two). Hence, for any point q in the ball $B_{\frac{1}{2}r_p}(p)$ the map $\exp_q$ is defined on a ball with center 0 and radius $\frac{1}{2}r_p$ in the tangent space M_q. Let I be the interval $I = (-\frac{1}{2}r_p, \frac{1}{2}r_p)$ in $\mathbb{R}$.
Because $\mathcal{D}_{\tau_{t1}}$ is involutive and of constant dimension $n-1$, this distribution has integral manifolds through every point (Frobenius' Theorem). Let N be the integral manifold of $\mathcal{D}_{\tau_{t1}}$ through p. Then N is a differentiable manifold and $N_q = \mathcal{D}_{\tau_{t1}}(q) \subset M_q$ for all $q \in N$. Let $(U_1, x_1 : U_1 \to V_1 \subset \mathbb{R}^{n-1})$ be a chart of N, such that $p \in U_1$. Let U be the interior in N of $U_1 \cap B_{\frac{1}{2}r_p}(p)$. Then, for any $q \in U$ the geodesic $\gamma_q(t) = \exp_q(tX(q))$ is well defined on the interval I. Let $V_1' = x_1(U)$. In that case V_1' is an open set in $\mathbb{R}^{n-1}$. Let $x_1(p) = a_0 \in V_1'$.
Now, we are able to define the map $\psi : V_1' \times I \to M$ by:

$$\psi(a,t) = \gamma_q(t) = \exp_q(tX(q)), \text{ if } x_1(q) = a \in V_1'. \tag{3.60}$$

Note that $\psi(a,0) = q$, if $x_1(q) = a$, hence in particular $\psi(a_0,0) = p$. The map ψ is C^∞,[6] and we shall calculate its derivative $\psi_{*|(a_0,0)} : \mathbb{R}^n \to M_p$ in the point $(a_0,0) \in V_1' \times I$. If $a : V_1' \to V_1'$ is the natural chart of V_1', i.e. the identity, and $t : I \to I$ the natural chart of I, then we have for any C^∞ function $f : M \to \mathbb{R}$:

$$\psi_{*|(a_0,0)}(\frac{\partial}{\partial a_i})f = \frac{\partial}{\partial a_i}(f \circ \exp_{x_1^{-1}(a)}(0))_{|a=a_0} = \frac{\partial}{\partial a_i}(f \circ x_1^{-1}(a))_{|a=a_0} = (\frac{\partial}{\partial x_{1\,i}})_p(f),$$

and

$$\psi_{*|(a_0,0)}(\frac{\partial}{\partial t}) = \exp_{p*|0}(X(p)) = X(p),$$

because the derivative of the normal map in the origin of the tangent space is the identity. This shows that the derivative of ψ in $(a_0,0)$ is bijective, because the tangent vectors $(\frac{\partial}{\partial x_{1\,1}})_p, \cdots, (\frac{\partial}{\partial x_{1\,n-1}})_p$ form a base of N_p, and $X(p)$ forms a base

[6] In chapter two we have recapitulated some basic facts about C^∞-maps between differentiable manifolds and their derivatives.

of $N_p^{\perp}$, so together they form a base of M_p. By means of the Implicit Function Theorem, we see that there exists an open set $V_\eta \subset V_1' \times I$ and an open set $U_\eta \subset M$, such that $\psi : V_\eta \to U_\eta$ is a diffeomorfism. Let $x_\eta : U_\eta \to V_\eta$ be the inverse of ψ. Then, (U_η, x_η) is a chart of M. We shall show that it is adapted to both distributions. Note simply that ψ is a pencil of geodesics, because the t-curves $\gamma_q(t) = \exp_q(tX(q))$ are geodesics parametrized, according to the arc length for all $q \in N$ (see chapter two). We have:

$$(\frac{\partial}{\partial x_{\eta\ n}})_{\gamma_q(t)} = \psi_{*|(a,t)}(\frac{\partial}{\partial t}) = \dot{\gamma}_q(t) = X(\gamma_q(t)), \text{ because } X \text{ is a geodesic tangent}$$

$$\text{field and } (\frac{\partial}{\partial x_{\eta\ i}})_{\gamma_q(t)} = \psi_{*|(a,t)}(\frac{\partial}{\partial a_i}) =: S_i \text{ (notation) if } 1 \leq i \leq n-1.$$

As was stated in chapter two, it is a well known fact that the inner product $\langle S_i, X\rangle_{\gamma_q(t)}$ of the vector fields S_i with the vector field X is constant along all geodesics $\gamma_q(t)$ of the pencil ψ. If $t = 0$ then $S_i(q) = (\frac{\partial}{\partial x_{1\ i}})_q$ so $\langle S_i, X\rangle_q = 0$ for all $q \in N$, and we conclude that $\langle S_i, X\rangle = 0$ everywhere on U_η.
X spans the distribution $\mathcal{D}_{\tau_{t2}}$, hence the chart (U_η, x_η) is adapted to this distribution, and the other coordinate base vector fields $\frac{\partial}{\partial x_{\eta\ i}} = S_i$ are perpendicular to X, so they span $\mathcal{D}_{\tau_{t1}}$. Consequently, the chart is also adapted to this distribution.
If we specify an orthonormal base of eigenvector fields $(F_{t,j})_{1\leq j\leq n}$ of V_t, which determine the Model, then necessarily $F_n = X$ everywhere or $F_n = -X$ everywhere, because M is connected. Writing out the linearization, using this basis on the chart (U_η, x_η), gives us immediately equations (3.58, 3.59). Finally, we know already that (3.59) is precise, which completes the proof. □

Example 3.10 (Continued): We continue with the situation of Example 3.10. According to proposition 2.41 of Nijmeijer and Van der Schaft, the distribution perpendicular to the distribution generated by the tangent field of all the geodesics through a point $q_{-1} \in \tilde{M}$ is involutive, because there is an integral manifold of it through each point in $M = U \setminus q_{-1}$. These integral manifolds are the spheres $S_r(q_{-1})$ of points with distance r to q_{-1}. It is also easy to make a direct construction of a chart that gives equations (3.58, 3.59). For instance, if $n = 2$ and e_1, e_2 is an orthonormal base of $M_{q_{-1}}$, then take $x_\eta = \psi^{-1}$, where for some $R > 0$

$$\psi : (-\pi, \pi) \times (0, R) \to M; \quad \psi(a,t) = \exp_{q_{-1}}(t(\cos(a)e_1 + \sin(a)e_2)). \quad \square$$

The following corollary 3.15 gives a sufficient condition for the existence of a similar type of chart, as constructed in the lemma 3.14 above, in case the distribution $\mathcal{D}_{\tau_{t2}}$ has higher dimension than one, or in case there are more than two eigenvalues. We shall use this corollary in chapter five and six in order to make an extension of the Geodesic Model: the Geodesic plus Noise Model.

Corollary 3.15 Consider a (General or Special) Model of a process $(q_t, d_t)_{t\in\mathbf{Z}_+}$ on a k-dimensional manifold M with direction spaces $\mathcal{L}_t$ and symmetric tensors V_t $(t \in \mathbf{Z}_+)$. Suppose:

1. that for some time t the r-dimensional distribution $\mathcal{D}_{\tau_{t1}}$, corresponding to the lowest eigenvalue τ_{t1} of V_t, is involutive,

2. that one eigen vector field $F_{t,r+1}$, corresponding to a higher eigenvalue τ_{t2} in $\mathcal{L}_t$, is a geodesic field and,

3. that the subspace $\tilde{\mathcal{L}}_t$ of $\mathcal{L}_t$ perpendicular to this field and to $\mathcal{D}_{\tau_{t1}}$ satisfies:

$$\nabla_X Y = 0 \text{ for all } X \in \mathcal{D}M \text{ and all } Y \in \tilde{\mathcal{L}}_t \text{ and}$$

$$\nabla_Y F_{t,r+1} = 0 \text{ for all } Y \in \tilde{\mathcal{L}}_t.$$

Then, there is a collection of charts $(U_\eta, x_\eta : U_\eta \to V_\eta)$ covering M,

- such that the linearization of the transition equation at time t,

$$q_{t+1} = \exp_{q_t}(d_t),$$

on this chart can be written as:
if $x_\eta(q_t) = (v_{t1}, \cdots, v_{tr}, w_{t1}, \cdots w_{t\ k-r})'$, and
if $x_\eta(q_{t+1}) = (v_{t+1\ 1}, \cdots, v_{t+1\ r}, w_{t+1\ 1}, \cdots w_{t+1\ \ k-r})'$ then,

$$v_{t+1} = v_t + \tilde{g}_\eta(v_t, w_t)\tau_{t1}\tilde{\lambda}_t \tag{3.61}$$

$$w_{t+1} = w_t + \begin{pmatrix} \tau_{t2} & \cdots & 0 \\ 0 & \ddots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & \tau_{tl} \end{pmatrix} \lambda_t \tag{3.62}$$

where $\begin{pmatrix} \tilde{\lambda}_t \\ \lambda_t \end{pmatrix}$ is unit variance white noise, and
$g(v,w) = \begin{pmatrix} \tilde{g}_\eta(v,w) & 0_{r\times k-r} \\ 0_{k-r\times r} & I_{r-k\times r-k} \end{pmatrix}$ satisfies $gg' = H_\eta^{-1}$, and

- such that there is no approximation error in (3.62).

Proof: Every element Y of the subspace $\tilde{\mathcal{L}}_t$ is a tangent field of geodesics, because of the third assumption . This space $\tilde{\mathcal{L}}_t$ is spanned by an orthonormal set of eigen vectors $F_{t,r+2}, \cdots, F_{t,k}$ of the tensor V_t. For $j \geq 0$, let $\mathcal{G}_j$ be the distribution of the vector fields perpendicular to $F_{t,r+2+j}, \cdots, F_{t,k}$. Then, $\mathcal{G}_j$ is spanned by $\mathcal{D}_{\tau_{t1}}$, $F_{t,r+1}$ and (if $j > 0$) the fields $F_{t,r+2}, \cdots, F_{t,r+1+j}$. We show that $\mathcal{G}_j$ is totally geodesic.
Let $X, Y \in \mathcal{G}_j$. For all $B \in \boldsymbol{span}\{F_{t,r+2+j}, \cdots, F_{t,k}\}$ we have: $\langle Y, B\rangle = 0$ everywhere, hence $X\langle Y, B\rangle = 0$. On the other hand we have: $X\langle Y, B\rangle = \langle \nabla_X Y, B\rangle$, because of the third assumption of this corollary. This shows that $\nabla_X Y \in \mathcal{G}_j$, hence $\mathcal{G}_j$ is totally geodesic. In particular, $\mathcal{G}_j$ is involutive. As it is also of constant dimension, there is an integral manifold $N_j(q)$ of $\mathcal{G}_j$ through any point $q \in M$. $F_{t\ r+1}$ is still a geodesic

tangent field on $N_0(q)$, and $\mathcal{D}_{\tau_{t1}}$ is involutive on $N_0(q)$. By lemma 3.14, there is a chart $(\tilde{U}_\eta, x_\eta : \tilde{U}_\eta \to \tilde{V}_\eta)$ of $N_0(q)$ around the point q, adapted to $\mathcal{D}_{\tau_{t1}}$, such that one of the coordinate vector fields is $F_{t\ r+1}$. Just as in the proof of the lemma we can show that there is an open interval $I \subset \mathbb{R}$ around 0 and a subset V' of $\tilde{V}_\eta$, such that the map $\psi : V' \times I \to M$, defined by:

$$\psi(a,s) = \exp_p(sF_{t,r+2}(p)), \text{ if } x_\eta(p) = a \in V'$$

has constant rank $r+2$ in the neighbourhood of $(a_0, 0)$ with $x_\eta(q) = a_0$. Again, we have: $\psi_{*|(a,t)}(\frac{\partial}{\partial t}) = F_{t,r+2}(\psi(a,t)))$, because $F_{t,r+2}$ is a tangent field of geodesics. As $N_1(q)$ is totally geodesic, we have in fact:

$$\psi : V' \times I \to N_1(q),$$

hence the inverse of ψ can be used as a chart of $N_1(q)$ around q. Denote the other derivatives by $C_i := \psi_{*|(a,t)}(\frac{\partial}{\partial a_i})$. Then, on $N_1(q)$, we have:

$$\nabla_{F_{t,r+2}} C_i = \nabla_{C_i} F_{t,r+2} = 0; \tag{3.63}$$

the first equality, because $C_i, F_{t,r+2}$ are coordinate vector fields on $N_1(q)$, and the second equality by assumption of the proposition. Also $F_{t,r+1}$ is a vector field on $N_1(q)$, and by assumption of the proposition we have $\nabla_{F_{t,r+2}} F_{t,r+1} = 0$. Let us write $C_{r+2} := F_{t,r+2}$. Then, equation (3.63) implies that on $N_1(q)$, $F_{t,r+2}\langle C_i, C_j\rangle = 0$ for all $1 \leq i,j \leq r+2$, while the assumption of the proposition also implies that $F_{t,r+2}\langle C_i, F_{t,r+1}\rangle = 0$. In particular, on $N_1(q)$, we have:

$$\langle C_i, F_{t,r+2}\rangle = 0 \text{ for all } 1 \leq i \leq r+1;$$

$$\langle C_i, F_{t,r+1}\rangle = 0 \quad (1 \leq i \leq r \text{ or } i = r+2); \quad \langle C_{r+1}, F_{t,r+1}\rangle = 1, \text{ and also}$$
$$\langle C_i, C_{r+1}\rangle = 0 \quad (1 \leq i \leq r); \quad \langle C_{r+1}, C_{r+1}\rangle = 1,$$

because on $N_0(q)$ $C_{r+1} = F_{t,r+1}$. From the last two statements, it is clear that $C_{r+1} = F_{t,r+1}$ anywhere on $N_1(q)$. Accordingly, we have constructed a chart on $N_1(q)$, such that the last two coordinate vector fields are $F_{t,r+1}$ and $F_{t,r+2}$, and, such that these are everywhere perpendicular to the other coordinate vector fields. By induction one can finish the proof. □

3.10 Convergence

In this section we present some considerations about running times and convergence. We denote the square root of the maximum eigenvalue of V_t by τ_t. So $\tau_{t,n} = \tau_t$. We say that a geodesic has finite length if its maximum interval of definition is bounded. For a geodesic with finite length, its length is equal to the length of the maximum interval, provided it is parametrized, according to the arc length. If, for $v \in M_p$, γ_v is the geodesic which satisfies

$$\gamma_v(0) = p \quad ; \dot{\gamma}_v(0) = v,$$

then we have $\gamma_v(t) = \gamma_{tv}(1)$, thus, if the length of the geodesic γ_v is less than R, then $\exp_p(tv)$ is not defined if $||tv||_p > R$.

Lemma 3.16 Suppose that there exists a positive number κ, such that

$$\tau_t > \kappa \quad \forall t \in \mathbf{Z}_+.$$

Suppose that all geodesics in M have a finite length, and that these lengths are uniformly bounded. Then, the running time is a.s. finite.

Proof:

If the maximal intervals of the geodesics, parametrized according to the arc length, are uniformly bounded, there is a number R, such that $\exp_q$ is not defined any more outside the ball $B_R(0)$ in M_q for all $q \in M$. Then, $\nu = \infty$ implies that $||d_t||_{q_t} \leq R$ for all $t \in \mathbf{Z}_+$. Then, also $|\theta_{t,n}\tau_t| = ||\theta_{t,n}\tau_{t,n}F_{t,n}|| < R$, hence

$$|\theta_{t,n}| < \frac{R}{\kappa} \text{ for all } t \in \mathbf{Z}_+.$$

Now, $(\theta_{t,n})_{t\in\mathbf{Z}}$ is a one-dimensional Gaussian stationary process with an everywhere positive spectral density. Just as we have seen above, there is a one-dimensional Gaussian white noise sequence $(\eta_t)_{t\in\mathbf{Z}}$, such that $(\eta_t)_{t\in\mathbf{Z}_+}$ is output of a BIBO-stable system with $(\theta_{t,n})_{t\in\mathbf{Z}_+}$ as input. Consequently, there is a number $\tilde{R}$, such that

$$|\eta_t| < \tilde{R} \text{ for all } t \in \mathbf{Z}_+.$$

Now, $P(|\eta_t| < \tilde{R}) = c < 1$ for all $t \in \mathbf{Z}_+$, because the η_t are Gaussian and identically distributed. As the η_t are also independent, we immediately have:

$$P(||d_t||_{q_t} < R \; \forall t \in \mathbf{Z}_+) \leq P(|\eta_t| < \tilde{R} \; \forall t \in \mathbf{Z}_+) = 0. \quad \square.$$

Remark 3.17 If all the geodesics of the manifold M' have finite maximal interval of definition, and $M \subset M'$ is an open subset with compact closure in M', then M is itself a Riemannian manifold. Its structure is inherited from M' and the maximal intervals of its geodesics, parametrized according to the arc length, are uniformly bounded. This can be concluded from the proof of the lemma 3.1 that shows that ν is a stopping time: the set $\tilde{W} = \{(q,w) \in TM' | \gamma(1,q,w) \text{ exists } \}$ is open. As any geodesic has a finite length, the function $q \mapsto \max\{||w|| \mid (q,w) \in \tilde{W}\}$ is well defined on M', continuous because $\tilde{W}$ is open, and hence bounded on the compact closure of M.
Another example of manifolds, where the geodesics have uniformly bounded maximal intervals, will be used in the next corollary: every point in a Riemmanian manifold is the center of a ball in which every two points can be joined by exactly one geodesic in that ball with a length equal to the distance of these two points (Helgason, Theorem I.9.9, see chapter two). The interior of such a ball is itself a Riemannian manifold with the property of the lemma.

Corollary 3.18 Suppose that there exists a positive number κ, such that $\tau_t > \kappa$ for all $t \in \mathbf{Z}_+$. Let M be an arbitrary Riemannian manifold. Then, the process $(q_t)_{t\in\mathbf{Z}_+}$, defined according to the General Model, does a.s. not converge.

Proof: As M is separable, we can cover M by a countable set $(B_i)_{i\in\mathbf{N}}$ of balls of the type Helgason, Theorem I.9.9 describes. If the process would converge, there would be a ball B_i that contains all but a finite number of the q_t. The lemma 3.16 above shows that this occurs with probability zero. □

This corollary shows that although, by linearizing the coordinates of q_t as in (3.35), one approximates the process by a martingale, the process itself does not inherit all the properties of an n-dimensional martingale. If the process is bounded it does e.g. not have to converge, as the example of a random walk on the circle (which obviously is a bounded Riemannian manifold) shows.

Lemma 3.19 Let M be an arbitrary Riemannian manifold. Let the process $(q_t, d_t)_{t\in\mathbf{Z}_+}$ be defined according to the General Model with running time ν. Let the $\mathcal{B}_t$-measurable variables h_{t+1} of equation (3.39) satisfy:

$$\boldsymbol{d}(q_t, h_{t+1}) \leq ||d_t||_{q_t}, \text{ a.s., if } \nu \leq t.$$

Let the square root of the maximum eigenvalue of the symmetric tensor field V_t be denoted by τ_t. Suppose that

$$\sum_{t=0}^{\infty} \tau_t < \infty.$$

Then $(q_t)_{t\in\mathbf{Z}_+}$ a.s. converges to a point of the manifold or to the boundary.

Proof: We say that a sequence $(p_n)_{n\in\mathbf{N}}$ in M is converging to the boundary, if for any compact subset K there is a number C, such that $p_n \notin K$ for all $n > C$. This is a logical definition, because there exists an increasing sequence of compact sets, the union of which is M (follows from local compactness and separability of M).
Now, first note that

$$\boldsymbol{E}\sum_{t=0}^{\infty} ||d_t||_{q_t} \leq \sum_{t=0}^{\infty} \sqrt{\boldsymbol{E}||d_t||^2_{q_t}} = \sum_{t=0}^{\infty} \sqrt{\boldsymbol{E}||\sum_{i=1}^{n} \theta_{t,i}\tau_{t,i}F_{t,i}||^2} =$$

$$= \sum_{t=0}^{\infty} \sqrt{\boldsymbol{E}\sum_{i=1}^{n} |\theta_{t,i}|^2\tau_{t,i}^2} \leq \sqrt{n \max_{i=1,\cdots,n}(\boldsymbol{Var}(\theta_{t,i}))} \sum_{t=0}^{\infty} \tau_t,$$

hence $\sum_{t=0}^{\infty} ||d_t||_{q_t}$ is a.s. finite. Next, introduce

$$x_t = \boldsymbol{E}^{\mathcal{B}_{t-1}}(\sum_{s=t}^{\infty} ||d_s||_{q_s}).$$

Then, we have:

$$\boldsymbol{E}^{\mathcal{B}_{t-1}} x_{t+1} = \boldsymbol{E}^{\mathcal{B}_{t-1}}(\boldsymbol{E}^{\mathcal{B}_t} \sum_{s=t+1}^{\infty} ||d_s||_{q_s}) = \boldsymbol{E}^{\mathcal{B}_{t-1}} \sum_{s=t+1}^{\infty} ||d_s||_{q_s} \leq x_t,$$

so $(x_t)_{t\in\mathbf{Z}_+}$ is a positive supermartingale, and consequently converges a.s. (see e.g. Neveu, Theorem II-2-9).
Let $p \in M$. The triangular inequality tells us that

$$d(q_{t+1},p) \le d(q_t,p) + d(q_{t+1},q_t) \le d(q_t,p) + ||d_t||_{q_t}.$$

Introduce the stochastic variables

$$z_t = d(q_t,p) + x_t.$$

Then,

$$E^{\mathcal{B}_{t-1}} z_{t+1} = E^{\mathcal{B}_{t-1}} d(q_{t+1},p) + E^{\mathcal{B}_{t-1}}(\sum_{s=t+1}^{\infty} ||d_s||_{q_s}) \le$$

$$\le d(q_t,p) + E^{\mathcal{B}_{t-1}}||d_t||_{q_t} + E^{\mathcal{B}_{t-1}}(\sum_{s=t+1}^{\infty} ||d_s||_{q_s}) = z_t,$$

hence, $(z_t)_{t\in\mathbf{Z}_+}$ is also a positive supermartingale, and it converges a.s., therefore

$$\lim_{t\uparrow\infty} d(q_t,p) \text{ exists a.s. for all } p \in M.$$

We are ready, if there is a point $p \in M$, such that $\lim_{t\uparrow\infty} d(q_t,p) = 0$. Now, assume that there is a compact set K and an infinite number of points of the sequence $(q_t)_{t\in\mathbf{Z}_+}$ in it. Then there is a converging subsequence, and we certainly have such a point p. On the other hand, if there is no point $p \in M$, such that $\lim_{t\uparrow\infty} d(q_t,p) = 0$, then any compact set contains only a finite number of points of the sequence $(q_t)_{t\in\mathbf{Z}_+}$, and the sequence converges to the boundary. □

Chapter 4

Asymptotics

The aim of this chapter is to answer questions about stability and asymptotic properties of an AR process with time-varying stochastically defined coefficients. These issues are important if one wants to analyze estimation procedures, e.g. for hyperparameters on which the conditional likelihood of such a process depends, given the stochastic process generating the coefficients. Unfortunately, it is hard to discuss stability and asymptotic properties in this context. Let us indicate the difficulty.
Let $(y_t)_{t\in\mathbf{Z}_+}$ be a univariate time-varying AR process satisfying

$$y_t = q_t'\phi_t + \sigma e_t \qquad (t \in \mathbf{Z}_+),$$

where $\phi_t = (y_{t-1}, \cdots, y_{t-n})'$, and $(e_t)_{t\in\mathbf{Z}}$ is Gaussian white noise with variance one. The coefficients $q_t = (a_{1\ t}, \cdots, a_{n\ t})'$ are vectors in $\mathbf{R}^n$. The set of all vectors $(a_1, \cdots, a_n)'$, such that the polynomial $x^n - a_1 x^{n-1} + \cdots - a_n$ is stable, is denoted by AR_n. This set is open and bounded in $\mathbf{R}^n$. Suppose that the transition equation for the coefficients q_t is of the form

$$q_{t+1} = q_t + cg(q_t)\lambda_t \quad (t \in \mathbf{Z}_+), \tag{4.1}$$

where $(\lambda_t)_{t\in\mathbf{Z}_+}$ is i.i.d. $\mathcal{N}(0, I_n)$, and g is a continuous matrix function, such that $g(q)$ is invertible for every $q \in AR_n$. For an arbitrary compact subset $K \subset AR_n$, we define the stopping time ν_K:

$$\nu_K = \inf\{t \in \mathbf{Z}_+ \mid q_t \notin K\}.$$

Then, it is easy to verify that [1]

$$P\{\nu_K = \infty\} = 0.$$

Moreover, if the coefficients q_t would remain in AR_n forever, then they would converge, because equation (4.1) shows that the coefficients are modelled as a martingale, and AR_n is bounded (application of theorem II-2-9, Neveu, 1975). Because they get out of any compact subset of AR_n in finite time, the only possibility is that they then a.s. converge to a point on the boundary of AR_n, and such a point corresponds to

[1] If, instead, the coefficient process follows a Special or General Model as defined in chapter three, we can make the same statement for any compact subset K of a neighbourhood in which all geodesics have finite length, according to Lemma (3.16).

an unstable polynomial. Hence, we face the problem that with equation (4.1) stability of the resulting time-varying infinite AR(n) process is very unlikely. Because of this, questions as stability and asymptotic analysis, just based on *one* infinite time-varying AR(n) process, look quite odd. If we are only interested in stable models of a process, then we are particularly interested in the part of the process until ν_K, provided the compact set K is large in AR_n. In order to perform asymptotic analysis, we have to look at a *sequence* of finite parts of such time-varying AR(n) processes. The first idea to get such a sequence is: sample a continuous time process with increasing frequency. In the next section, we shall explain why we think that this is *not* a good idea. It was Dahlhaus who, in the author's opinion, was the first to recognise that it is necessary to consider a sequence of time-varying AR processes to carry out meaningful asymptotic analysis. He also had the view that such a sequence does not necessarily originate from ever more frequent sampling of *one* continuous time process. In the next section, we shall give a fictitious example to explain our understanding of his ideas. This will be done with a model with deterministic coefficients. In section (4.3), we introduce Dahlhaus' notion of time-varying spectral density for a sequence of time-varying AR processes with *deterministic* coefficients. We shall link it with the concept of stability in a similar way as in the time-invariant case. This concept will be introduced in section (4.2). In sections (4.4, 4.5) and (4.6), we introduce the time-varying spectral density of Dahlhaus for a sequence of time-varying AR processes with *stochastic* coefficients. This spectral density will be conditional on the stochastic process generating the coefficients.

4.1 The asymptotics of Dahlhaus

The probability density of a discrete time zero mean Gaussian stationary process is completely determined by its autocovariance function, and hence, by its spectrum. One can say that all the information about the process is contained in the spectrum. Ergodic theorems show that it is possible to know this spectrum more and more accurately from a single realization by taking more datapoints into account. Asymptotic theory for Gaussian stationary processes can almost completely be based on this spectrum and ergodic theorems.
In general, time-varying AR processes are not stationary, although there have been some reports of very special cases of stationary time-varying AR processes (with random coefficients, Pourahmadi, 1988).
Yet, with the idea in mind to have a basket for collecting all the information, one can search for the possibility to construct a *time-varying* spectrum. Such a spectrum should contain all the relevant information, hence we expect that

- such a time-varying spectrum facilitates to find workable expressions for e.g. the probability density, Kullback-Leibler divergence etc. for time-varying AR processes (conditional on the stochastic process generating the coefficients).

- Furthermore, the time-varying spectrum should help us in constructing interesting filters.

- It would be nice if there is some direct relation between the time-varying spectrum and the coefficients of the process. (This is one of the reasons that we did not adopt the evolutionary spectral theory, cited in the Introduction (chapter one). An evolutionary spectral density of a time-varying AR process has a complicated relation with the coefficients, see e.g. Mélard and Herteleer-de Schutter (1988)).

- Finally, a theory of time-varying AR processes with time-varying spectrum should contain the theory of time-invariant stationary AR processes with the ordinary spectrum as a special case.

One can't hope for ergodic theorems for a time-varying spectrum in general: how could we expect that an increase in the number of datapoints between e.g. time 1000 and time 2000 would help us in knowing more accurately the time-varying spectrum between time 100 and time 200? This will only be possible if the spectrum in early times has much in common with the spectrum in later times. If this is not the case, then we need more than only the datapoints of *one* time-varying AR process in order to obtain precise knowledge of such a spectrum.
The first idea could be to intensify the sampling between time 100 and 200. This opinion stands for thinking of the time-varying AR process as a discretization of a continuous-time process that has a time-varying spectrum. By more intensive sampling of this process, we would then get better knowledge of this time-varying spectrum. However, we think that this thought is wrong. Let us explain our view.
Firstly, it is not always possible to see a discrete-time *time-invariant* AR(n) process as a discretization of a continuous-time time-invariant AR(n) process. This is possible for an AR(1) process with a positive coefficient smaller than one. Such a process can be embedded in a Ornstein-Uhlenbeck Velocity process. However, for an AR(1) process with negative coefficient this is already not possible anymore (see e.g. Chan and Tong, 1987). Necessary and sufficient conditions for embedding a discrete-time time-invariant AR(2) process in a continuous-time process of the same kind are given in He and Wang (1989). These conditions imply *far more* than just stationarity.
Now, we will try to explain the difficulties which can arise in the time-varying case.
Let $(y_t)_{t\in\mathbf{Z}}$, $y_t : \Omega \to \mathbf{R}$ be an autoregressive stationary process with deterministic time-invariant coefficients and with spectral density $f(x) = \sum_{k=-\infty}^{\infty} f_k x^k$, $(x \in C, |x| = 1)$, such that $\sum_{k=-\infty}^{\infty} |f_k| < \infty$. Let $(z_t)_{t\in\mathbf{Z}}$ be the process that we obtain by sampling the process $(y_t)_{t\in\mathbf{Z}}$ only at time points that are multiples of a certain integer p:

$$z_t = y_{pt} \quad \forall t \in \mathbf{Z}.$$

Then, also $(z_t)_{t\in\mathbf{Z}}$ is a stationary process with spectral density $g(x) = \sum_{k=-\infty}^{\infty} g_k x^k$, $(x \in C, |x| = 1)$, and it is easily recognised that we have:

$$g_k = f_{pk}, \tag{4.2}$$

because

$$\boldsymbol{E} z_{t+k} z_t = \boldsymbol{E} y_{pt+pk} y_{pt}.$$

Now, let $(y_t)_{t\in\mathbf{Z}}$ be a time-varying AR process with deterministic coefficients, and suppose that we would have been able to define a time-varying spectral density

$$f(t,x) = \sum_{k=-\infty}^{\infty} f_{t,k}x^k \qquad (x \in \mathcal{C},\ |x| = 1).$$

I seems reasonable to assume that this density can be decomposed as $f(t,x) = |A(t,x)|^2$, where $A(t,.)$ is a Hardy function on the unit disk. We shall require that $A(t,.)$ is maximal, i.e. if $f(t,x) = |B(t,x)|^2 \quad (x \in \mathcal{C}, |x| = 1)$ and $B(t,.)$ is a Hardy function, then $|A(t,0)|^2 \geq |B(t,0)|^2$. We shall also require that $A(t,0) \in \mathbf{R}_+$, which implies that the decomposition is unique[2]. A "reasonable" definition of time-varying spectral density should satisfy:

$$\boldsymbol{E}y_{t+k}y_t = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\phi k}A(t+k,e^{i\phi})\overline{A(t,e^{i\phi})}d\phi. \tag{4.3}$$

This, because we define time-varying spectra in order to simplify expressions for the likelihood of the process $(y_t)_{0\leq t\leq T}$, Kullback-Leibler divergence, and so on. If the process is Gaussian, the likelihood is completely determined by the autocovariances of the process. At least, one should be able to express those, using the time-varying spectrum.
If again $(z_t)_{t\in\mathbf{Z}}$ is the less frequently sampled process as above, we would get:

$$\boldsymbol{E}z_{t+k}z_t = \boldsymbol{E}y_{pt+pk}y_{pt} = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\phi pk}A(pt+pk,e^{i\phi})\overline{A(pt,e^{i\phi})}d\phi. \tag{4.4}$$

We can write

$$A(t,x) = \sum_{l=0}^{\infty} a_{t,l}x^l\ ,\ (x \in \mathcal{C}, |x| = 1) \text{ so } f_{t,k} = \sum_{l=0\vee -k}^{\infty} a_{t,l}\overline{a_{t,l+k}} \text{ and}$$

$$\boldsymbol{E}y_{t+k}y_t = \sum_{l=0\vee -k}^{\infty} a_{t+k,l}\overline{a_{t,l+k}}\ ,\ \text{so } \boldsymbol{E}z_{t+k}z_t = \sum_{l=0\vee -pk}^{\infty} a_{pt+pk,l}\overline{a_{pt,l+pk}}.$$

This shows that a time-varying spectral density $g(t,x)$ for the process $(z_t)_{t\in\mathbf{Z}}$ could be:

$$g(t,x) = \sum_{k=-\infty}^{\infty} g_{t,k}x^k, \text{ where } g_{t,k} = \sum_{l=0\vee -pk}^{\infty} a_{pt,l}\overline{a_{pt,l+pk}} = f_{pt,pk},$$

but *not* $g(t,x) = f(pt,x)$, because under the same "reasonable" requirements we would automatically have:

$$\boldsymbol{E}z_{t+k}z_t = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\phi k}A(pt+pk,e^{i\phi})\overline{A(pt,e^{i\phi})}d\phi$$

in contradiction with (4.4).

Conclusion:

[2]A necessary and sufficient condition for the existence of such a decomposition is that

$$\int_{-\pi}^{\pi} \log f(t,e^{i\phi})d\phi > -\infty.$$

This, and all we have stated in the last paragraph, can be found in Rozanov, pages 57-64.

> It is not possible to give a "reasonable" definition of a time-varying spectral density $f(t,x)$ for the process $(y_t)_{t\in\mathbf{Z}}$, such that the time-varying spectral density for the process $(z_t)_{t\in\mathbf{Z}}$, $z_t = y_{pt}$ would be $f(pt,x)$.

This means that it is hard to construct an asymptotic theory, based on just ever more frequent sampling of the same process, in order to obtain asymptotically accurate knowledge of the time-varying spectrum. If we go from $(z_t)_{t\in\mathbf{Z}}$ to $(y_t)_{t\in\mathbf{Z}}$, then we do not only have to deal with an increasing number of time points of the spectral density, (which perhaps could be dealt with by continuity in t arguments for $f(t,x)$), but also with an increasing number of spectral coefficients, like $f_{pt,pk+1},\cdots f_{pt,pk+p-1}$. The number of spectral coefficients, to be estimated, is therefore multiplied by p^2, whereas the number of new data is only p times the former number of data.
Even when we do not seek refuge in time-varying spectra, the same phenomenon can be observed by taking a look at the autocovariances. If the process is time-varying, then sampling more frequently, i.e. going from process $(z_t)_{t\in\mathbf{Z}}$ to $(y_t)_{t\in\mathbf{Z}}$, would multiply the number of autocovariances to be estimated by p^2, whereas the number of new data is only p times the former number of data. Hence, it is difficult to formulate an asymptotic theory what so ever if it is based only on one ever more frequently sampled process.
Dahlhaus thought of another possibility: in order to get more accurate knowledge of a time-varying spectrum one should (abstractly) think of a *sequence* of processes, using the same time-varying spectrum ever more intensively. For instance, an ordinary stationary infinite process can be seen as a sequence of finite processes. The tenth process contains the ten datapoints from time 1 until time 10 and process number hundred contains the hundred datapoints from time 1 until time 100. The autocovariances in every process of the sequence are determined by a common spectrum. But the sequence of processes can be of another kind too. There are e.g. processes in the sequence with 100 datapoints between time 100 and 200, but also with 1000 datapoints between these times. The processes have autocovariances that, with increasing number of datapoints, are more and more determined by the same time-varying spectrum. However, we don't suppose any other relation between the processes in the sequence. In order to make this idea less abstract, we composed a fictitious example of such a sequence of processes.

4.1.1 A time-varying doubly indexed AR(1) process

Completely fictitious. January 1994. Person 0 should pay a certain amount of money to person 2. The money should be transferred to person 2 within one year. She does not know person 2, but she does know person 1, who knows person 2. So she sends the money immediately to person 1, who waits half the time (half a year) after which he transfers the money to person 2. Both use the same Bank for the transferences. The Bank charges a certain promillage of the amount of money for transferring it, and charges also some rather accidental fees like telephone or telex costs. Also person 1 can charge telephone costs etc., if he didn't like the job or add something to the amount, if he liked it. As the Bank is competing with other banks, the charged promillage may vary from day to day. Let $y_{i,2}$ be the amount of money sent ,$i < 2$, or received ,$i = 2$, by person i.

Then, we have:

$$y_{1,2} = a(0)y_{0,2} + e_{0,2} \tag{4.5}$$
$$y_{2,2} = a(\tfrac{1}{2})y_{1,2} + e_{1,2} \tag{4.6}$$

In these equations is $1000 \times (1 - a(p))$ the promillage that the Bank charges on the day that fraction p of the year 1993 has passed ($p = 0$ corresponds to January 1; $p = \frac{1}{2}$ corresponds to July 1). The noise term $e_{i,2}$ stands for the accidental other subtractions or additions made by Bank or person i.
We shall call such a situation a *one-year chain of payments*. Of course, such a "one-year chain of payments" may consist of more than two links. For instance, if it consists of four links, and if $y_{i,4}$ is the amount sent ($i < 4$) or received ($i = 4$) by person i, we have:

$$y_{1,4} = a(0)y_{0,4} + e_{0,4} \tag{4.7}$$
$$y_{2,4} = a(\tfrac{1}{4})y_{1,4} + e_{1,4} \tag{4.8}$$
$$y_{3,4} = a(\tfrac{1}{2})y_{2,4} + e_{2,4} \tag{4.9}$$
$$y_{4,4} = a(\tfrac{3}{4})y_{3,4} + e_{3,4} \tag{4.10}$$
$$\tag{4.11}$$

If we consider the extra payment term $e_{i,4}$ as stochastic, i.i.d. normally distributed quantities (i.e. white noise), then the four-link one-year chain of payments $(y_{i,4})_{i\in\{0,1,2,3,4\}}$ becomes a time varying AR(1) process.
Now, suppose that the Bank becomes bankrupt in 1995. There is almost no information anymore about the promillages charged on every day in the year 1994. A committee of inquiry has only the data $y_{t,T}$ of some chain(s) of payments. Can the committee reconstruct the day promillage function? (The only fact known is that the day promillage function belongs to a certain parametrized set of time functions). Dahlhaus' answer to this problem would be: YES provided the promillage function $a : [0,1] \to \mathbf{R}$ is smooth, and we have a chain of payments with a sufficient number of links at our disposal. To be more precise: we can determine the parameters characterizing the smooth promillage function asymptotically accurately if we have a series of chains of payments available with a number of links increasing to infinity. To obtain a certain degree of accuracy, it is only necessary to analyze one chain out of this series with a sufficient number of links.

In Dahlhaus' idea the coefficient $a(t)$ will be deterministic, but it is not difficult to extend the idea to stochastic coefficients. In this example, this would mean that the promillage function follows a stochastic mechanism. A committee, trying to reconstruct this mechanism in the year 1994, should not only look at chains of payments via one bank, but also via other banks, of which the promillage function is thought to follow the same stochastic mechanism.

4.2 Stability of a sequence of time-varying systems

In this section, the main object of study will be a sequence of systems $(\Sigma_T)_{T\in\mathbf{N}}$:

$$\Sigma_T : \left\{ \begin{array}{lcl} X_{k+1,T} & = & A_{k,T}X_{k,T} + B_{k,T}u_{k,T} \\ y_{k,T} & = & C_{k,T}X_{k,T} \end{array} \right\} \quad (0 \le k \le T) \tag{4.12}$$

Here, the $A_{k,T}$'s are linear maps from $\mathbb{R}^n$ to $\mathbb{R}^n$,the $B_{k,T}$'s are linear maps from $\mathbb{R}^m$ to $\mathbb{R}^n$, and the $C_{k,T}$'s are linear maps from $\mathbb{R}^n$ to $\mathbb{R}^p$. The states $X_{k,T}$ are vectors in $\mathbb{R}^n$, the inputs $u_{k,T}$ are vectors in $\mathbb{R}^m$, and the outputs $y_{k,T}$ are vectors in $\mathbb{R}^p$. By "initial states" we mean the vectors $X_{0,T}$ $(T \in \mathbb{N})$.
Note that of every element Σ_T out of this sequence of systems $(\Sigma_T)_{T\in\mathbb{N}}$ only a finite number of path points is defined; only the points $(u'_{k,T}, X'_{k,T}, y'_{k,T})$ with $0 \leq k \leq T+1$. Because of this, a concept of *stability* does not make much sense for any particular element of the sequence . Yet, it is interesting to introduce such a concept for the sequence as a whole.

Definition. We shall call the sequence of systems $(\Sigma_T)_{T\in\mathbb{N}}$ defined above (4.12) *BIBO (Bounded Input, Bounded Output) stable* if for every positive number M there exists a positive number N, such that for every input u, satisfying $||u_{k,T}|| \leq M$ for all k, $0 \leq k \leq T$ for all $T \in \mathbb{N}$, we have for the corresponding output with zero initial state that $||y_{k,T}|| \leq N$ for all k, $0 \leq k \leq T$ for all $T \in \mathbb{N}$. □

This definition can very well be compared with the concept of *BIBO stability* in the single system case. It is, hence, not amazing that we have a same type of criterion for testing BIBO stability for $(\Sigma_T)_{T\in\mathbb{N}}$ as for testing BIBO stability for a single system Σ:

Lemma 4.1 The sequence of systems $(\Sigma_T)_{T\in\mathbb{N}}$ is BIBO stable if and only if there exists a number G, such that

$$\sum_{j=1}^{k} ||C_{k,T}A_{k-1,T}\cdots A_{j,T}B_{j,T}|| < G \text{ for all } k,\ 0 \leq k \leq T \text{ for all } T \in \mathbb{N}.$$

Proof: If the initial state is zero, the output y can be written as

$$y_{k,T} = \sum_{j=1}^{k} C_{k,T}A_{k-1,T}\cdots A_{j,T}B_{j,T}u_{j,T}, \tag{4.13}$$

or, more briefly,

$$y_{k,T} = \sum_{j=1}^{k} K_{k,j,T}u_{j,T},$$

where

$$K_{k,j,T} = C_{k,T}A_{k-1,T}\cdots A_{j,T}B_{j,T} \in \mathbb{R}^{p\times m}.$$

Suppose that the sequence of systems is BIBO stable. Let M, N be as in the definition of BIBO stability. Choose arbitrary $T \in \mathbb{N}$; $k, 0 \leq k \leq T$; $i, 1 \leq i \leq m$ and $s, 1 \leq s \leq p$. Define the following input u:

$$(u_{j,T})_i = M\text{sign}((K_{k,j,T})_{s,i}) \text{ if } 1 \leq j \leq k;$$

$$(u_{r,S})_v = 0 \text{ for all other indices.}$$

The input satisfies $||u_{r,S}|| \leq M$ for all r, S. Hence, the corresponding output $y_{r,S}$ with initial state zero should satisfy $||y_{r,S}|| \leq N$. For this output, we have:

$$(y_{k,T})_s = M \sum_{j=1}^{k} |(K_{k,j,T})_{s,i}|.$$

Hence, $\sum_{j=1}^{k} |(K_{k,j,T})_{s,i}| \leq \frac{N}{M}$ for all k, T, s, i. This implies the criterion.
Suppose that the criterion is satisfied. Let M be as in the definition above. Equation (4.13) immediately implies the definition: take $N = GM$. □

In the case of a time invariant single system

$$\Sigma : \left\{ \begin{array}{lcl} X_{k+1} & = & AX_k + Bu_k \\ y_k & = & CX_k \end{array} \right\},$$

we have also another criterion to establish BIBO-stability: it is sufficient that the matrix A has all its eigenvalues in the interior of the unit disk. This is even necessary if the system is controllable and observable. We shall show that we have the same type of criterion for sequences of systems of type (4.12) if the system matrices are slowly varying in a certain way.

Definition. The system matrices of a sequence of systems with increasing number of path points (4.12) are said to vary *equislowly* (with that number of path points), if there exists a number C, such that

$$\left\{ \begin{array}{lcl} ||A_{k+1,T} - A_{k,T}|| & < & \frac{C}{T} \\ ||B_{k+1,T} - B_{k,T}|| & < & \frac{C}{T} \\ ||C_{k+1,T} - C_{k,T}|| & < & \frac{C}{T} \end{array} \right\} \text{ for all } k,\ 0 \leq k < T \text{ and for all } T \in \mathbb{N}.$$

Let us suppose that the system matrices are equislowly varying, and moreover, are uniformly bounded. We consider the polygons $(C, A, B)_T$ through the matrices $(C_{k,T}, A_{k,T}, B_{k,T})_{0 \leq k \leq T}$, defined on the segment $[0, 1]$ as follows:

$$(C, A, B)_T(0) = (C, A, B)_{0,T}.$$

$$\text{If } \frac{k}{T} < u \leq \frac{k+1}{T}, \text{ then}$$

$$(C, A, B)_T(u) = (C_{k,T}, A_{k,T}, B_{k,T}) +$$

$$+ (Tu - k)((C_{k+1,T}, A_{k+1,T}, B_{k+1,T}) - (C_{k,T}, A_{k,T}, B_{k,T})). \qquad (4.14)$$

The collection of these polygons forms an equicontinuous set of curves on $[0, 1]$ [3]. According to the theorem of Arzela-Ascoli, this set of polygons has a compact closure under the supremum norm. Of course, also its subset $\mathcal{D}$ of limit curves is compact and equicontinuous. The union D of the images of these limit curves is compact as well, because the function

$$[0, 1] \times \mathcal{D} \to \mathcal{L}(\mathbb{R}^n \to \mathbb{R}^p) \times \mathcal{L}(\mathbb{R}^n) \times \mathcal{L}(\mathbb{R}^m \to \mathbb{R}^n);$$

[3] See chapter two for equicontinuity, equiLipschitz-property, and the theorem of Arzela Ascoli.

$$(x, f) \mapsto f(x)$$

is continuous. Next, we define the projection D_2 of D on $\mathcal{L}(\mathbb{R}^n)$:

$$D_2 = \{A \in \mathcal{L}(\mathbb{R}^n) | \ \exists C, B \ [(C, A, B) \in D]\}. \tag{4.15}$$

Then, D_2 is compact too.
The nice point of this construction is that many system theoretical concepts for single systems with time-invariant system matrices can be transferred to completely similar concepts for sequences of systems with increasing number of path points, and equislowly varying, uniformly bounded system matrices. These will have similar properties. We shall only demonstrate this similarity in the case of BIBO stability. In the same way this can be done in the case of controllability and observability.
Let U be the set of asymptotically stable matrices of size $n \times n$:

$$U = \{A \in \mathcal{L}(\mathbb{R}^n) | \lambda \text{ eigenvalue of } A \rightarrow |\lambda| < 1\}. \tag{4.16}$$

It is well known that U is open. Also well known is the following lemma 4.2:

Lemma 4.2 $A \in U$, if and only if there exists an operator norm, such that its value in A is smaller than one.

(*Proof:* If $A \in U$, then simply take a number $\mu > 1$, such that μA is still in U, and define the norm on $\mathbb{R}^n$: $\nu_A(x) = \sum_{j=0}^{\infty} ||(\mu A)^j(x)||$. This form is well defined, because $\lim_{n\to\infty} ||A^n||^{\frac{1}{n}} = \max(|\lambda|)$, λ eigenvalue of A. The rest is trivial.) □

Corollary 4.3 On any compact subset of U, the form $\sum_{j=0}^{\infty} ||A^j||$ is bounded, and the sequence of power functions $A \mapsto A^k$ $(k \in \mathbb{N})$ is equiLipschitz.

Proof: If $A \in U$, there is a compact neighbourhood of A in which all the matrices B have values less than one under the same operator norm. By lemma 4.2 and the equivalence of norms, the first statement immediately follows. To prove the second one, we make the following five observations.
1) If a function $f : V \subset \mathbb{R}^p \rightarrow \mathbb{R}^q$ is continuously differentiable, then for any points x,y, such that $(1 - \lambda)x + \lambda y \in V$ for all $\lambda \in [0, 1]$ it follows that:

$$||f(y) - f(x)|| \leq \sup_{\lambda\in[0,1]} ||Df_{(1-\lambda)x+\lambda y}|| \ ||x - y||.$$

2) The function $f : \mathbb{R}^{n\times n} \rightarrow \mathbb{R}^{n\times n}$, defined by $f(A) = A^k$, is continuous-differentiable, and $Df_A(H) = \sum_{j=0}^{k-1} A^{k-1-j} H A^j$.
3) As U is open, for any compact subset K of U, there is an $\varepsilon > 0$, such that $A + H \in U$ for all $A \in K$, and H satisfying $||H|| \leq \varepsilon$.
4) The set K^1 of all matrices $A + H$, where $A \in K$ and $||H|| \leq \varepsilon$ is again compact.
5) According to the first statement, $\sup_{B\in K^1}(\sum_{j=0}^{\infty} ||B^j||) < \infty$.
Combining these five observations, we get the following inequality for $A \in K$, and $||H|| \leq \varepsilon$:

$$||(A + H)^k - A^k|| \leq ||H|| \sup_{\lambda\in[0,1]} \sum_{j=0}^{k-1} ||(A + \lambda H)^{k-1-j}|| \ ||(A + \lambda H)^j|| \leq$$

$$\leq ||H|| \sup_{\lambda\in[0,1]} \sum_{k=1}^{\infty} \sum_{j=0}^{k-1} ||(A+\lambda H)^{k-1-j}||\ ||(A+\lambda H)^j|| =$$

$$= ||H|| \sup_{\lambda\in[0,1]} (\sum_{j=0}^{\infty} ||(A+\lambda H)^j||)^2 \leq ||H|| \sup_{B\in K^1} (\sum_{j=0}^{\infty} ||B^j||)^2.$$

This shows that the set of power functions $A \mapsto A^k$ on K ($k \in \mathbb{N}$) is equiLipschitz. □

Now, we are able to formulate another criterion to establish BIBO stability.

Proposition 4.4 Assume that the sequence of systems $(\Sigma_T)_{T\in\mathbb{N}}$, as defined in (4.12), has uniformly bounded system matrices and is equislowly varying with the increasing number of path points. Define the sets D_2 and U as in equations (4.15, 4.16). Then:

1. if $D_2 \subset U$, the sequence is BIBO stable;

2. if the sequence is BIBO stable, then for all triples $(C, A, B) \in D$, the invariant single system (C, A, B) is BIBO stable.

Proof:

1. Because of uniform convergence, the compactness of D_2 and the openness of U, there exists a compact set $K \subset U$ and a number S, such that for all $T > S$ and $k, 0 \leq k \leq T$, $A_{k,T} \in K$. We shall prove that the criterion in the first lemma 4.1 of this section is satisfied by using a Gronwall-type of iterative inequalities. Thus, we shall prove that

$$\sum_{j=1}^{k+1} ||A_{k,T} \cdots A_{j,T}|| = \sum_{j=1}^{k} ||A_{k,T} \cdots A_{j,T}|| + ||I||$$

is bounded uniformly in k, T.
Note that by equislow variance of the system matrices, and the equiLipschitz property of the power functions on K , there exists a number C, such that

$$||A^q_{j+1,T} - A^q_{j,T}|| < \varepsilon_T \quad \forall j, T > S, q \in \mathbb{N},$$

where $\varepsilon_T = \frac{C}{T}$.
Using this inequality we have:

$$||A_{k,T} \cdots A_{j,T}|| \leq ||A_{k,T} \cdots A_{j+1,T}||||A_{j+1,T} - A_{j,T}|| + ||A_{k,T} \cdots A^2_{j+1,T}|| \leq$$

$$\leq \varepsilon_T||A_{k,T} \cdots A_{j+1,T}|| + ||A_{k,T} \cdots A_{j+2,T}||||A^2_{j+2,T} - A^2_{j+1,T}|| + ||A_{k,T} \cdots A^3_{j+2,T}|| \leq$$

$$\leq \varepsilon_T \sum_{i=j+1}^{k} ||A_{k,T} \cdots A_{i,T}|| + ||A^{k-j+1}_{k,T}||.$$

Conclusion:

$$\sum_{j=1}^{k}||A_{k,T}\cdots A_{j,T}|| \leq \varepsilon_T \sum_{j=1}^{k-1}\sum_{i_1=j+1}^{k}||A_{k,T}\cdots A_{i_1,T}|| + \sum_{j=1}^{k}||A_{k,T}^{k+1-j}||. \qquad (4.17)$$

Next, by iterating (4.17) we get:

$$\sum_{j=1}^{k}||A_{k,T}\cdots A_{j,T}|| \leq \varepsilon_T^2 \sum_{j=1}^{k-1}\sum_{i_1=j+1}^{k-1}\sum_{i_2=i_1+1}^{k}||A_{k,T}\cdots A_{i_2,T}|| +$$

$$+\varepsilon_T \sum_{j=1}^{k-1}\sum_{i_1=j+1}^{k}||A_{k,T}^{k+1-i_1}|| + \sum_{j=1}^{k}||A_{k,T}^{k+1-j}||,$$

hence,

$$\sum_{j=1}^{k}||A_{k,T}\cdots A_{j,T}|| \leq \sum_{s=0}^{k-1}\varepsilon_T^s \sum_{j=1}^{k-1}\sum_{i_1=j+1}^{k-1}\cdots\sum_{i_{s-1}=i_{s-2}+1}^{k-1}\sum_{i_s=i_{s-1}+1}^{k}||A_{k,T}^{k+1-i_s}||. \qquad (4.18)$$

Rearranging order, we can rewrite the coefficients in this power series of ε_T (we substitute $r_l = k+1-i_l$; $r_0 = k+1-j$):

$$\sum_{j=1}^{k-1}\sum_{i_1=j+1}^{k-1}\cdots\sum_{i_{s-1}=i_{s-2}+1}^{k-1}\sum_{i_s=i_{s-1}+1}^{k}||A_{k,T}^{k+1-i_s}|| = \sum_{r_0=2}^{k}\sum_{r_1=2}^{r_0-1}\cdots\sum_{r_s=1}^{r_{s-1}-1}||A_{k,T}^{r_s}|| =$$

$$= \sum_{r_s=1}^{k}\Big(\sum_{r_{s-1}=r_s+1}^{k}\cdots\sum_{r_0=r_1+1}^{k}\Big)||A_{k,T}^{r_s}||.$$

Now, the expression between parentheses satisfies:

$$\sum_{r_{s-1}=r_s+1}^{k}\cdots\sum_{r_0=r_1+1}^{k}1 = \sum_{r_0,\cdots r_{s-1};r_s<r_{s-1}<\cdots r_0<k}1 \leq \frac{T^s}{s!}.$$

So, according to corollary 4.3, there exists a number B, such that

$$\sum_{r_s=1}^{k}||A_{k,T}^{r_s}|| \leq B \quad \forall k, T > S.$$

Putting all this in the inequality (4.18) yields:

$$\sum_{j=1}^{k}||A_{k,T}\cdots A_{j,T}|| \leq \sum_{s=o}^{k-1}\frac{\varepsilon_T^s T^s}{s!}\sum_{r_s=1}^{k}||A_{k,T}^{r_s}|| \leq e^C B. \qquad (4.19)$$

This proves part 1 of the proposition 4.4.

2. If the sequence is BIBO stable, then according to the first lemma 4.1 of this section, there exists a number G, such that

$$\sum_{i=1}^{k} ||C_{j+k,T}A_{j+k,T}\cdots A_{j+i,T}B_{j+i,T}|| \leq G \quad \text{for all } j \text{ and } k, \text{ such that } j+k \leq T,$$

and for all T. Take a fixed number k. Consider the set V of $2k+1$-tuples of matrices

$$(C_{j+k,T}, A_{j+k,T}, B_{j+k,T}, A_{j+k-1,T}, B_{j+k-1,T}, \cdots, A_{j+1,T}, B_{j+1,T}) \in$$

$$\in X := \mathcal{L}(\mathbb{R}^n \to \mathbb{R}^p) \times (\mathcal{L}(\mathbb{R}^n) \times \mathcal{L}(\mathbb{R}^m \to \mathbb{R}^p))^k.$$

It is a precompact set; the set of its limit points being

$$\{(C,A,B,A,B,\cdots,A,B)|(C,A,B) \in D)\}.$$

The map $X \to \mathbb{R}$,

$$(C, A_k, B_k, \cdots, A_1, B_1) \mapsto \sum_{i=1}^{k} ||CA_kA_{k-1}\cdots A_iB_i||$$

is continuous, and bounded by G on the set V. Hence, on the set of the limit points of V, the map is bounded by G as well:

$$\sum_{i=1}^{k} ||CA^iB|| < G \quad \forall (C,A,B) \in D.$$

As this is valid for all $k \in \mathbb{N}$, the invariant single systems with system matrices $(C,A,B) \in D$ are all BIBO stable. □

Remark 4.5 It is not sufficient for BIBO stability of the sequence of systems that all $(C,A,B) \in D$ are BIBO stable. The construction of a counter example goes as follows. The set of matrices (C,A,B), such that (C,A) is observable and (A,B) is controllable, forms an open subset of $\mathcal{L}(\mathbb{R}^n \to \mathbb{R}^p) \times \mathcal{L}(\mathbb{R}^n) \times \mathcal{L}(\mathbb{R}^m \to \mathbb{R}^n)$, its complement having an empty interior. (Look at the sum of subdeterminants of the matrix $\begin{pmatrix} B & AB & \cdots & A^{n-1}B \end{pmatrix}$ to recognise this). Fix a matrix C and B. Take an unstable matrix A_∞ (one eigenvalue having modulus one), such that the smallest A_∞-invariant subspace of $\mathbb{R}^n$, containing ImB, $< A_\infty|ImB >$, has dimension $n-1$, $(A_\infty)_{|<A_\infty|ImB>}$ is stable, and (C, A_∞) is observable. Take the half line in $\mathcal{L}(\mathbb{R}^n)$, beginning in 0 through A_∞. It starts in U and reaches the border of U in the point A_∞. In A_∞, we still have that $\sum_{i=0}^{\infty} ||CA_\infty^iB|| < \infty$. Now, we can approximate this half line by other half lines starting at 0 through matrices A_k, such that (C, A_k, B) are observable and controllable. Consequently, if the half line through A_k leaves U in μ_kA_k, then $\sum_{i=0}^{\infty} |\lambda^i|\ ||CA_k^iB|| = \infty$ if $\lambda = \mu_k$, and $\sum_{i=0}^{\infty} |\lambda^i|\ ||CA_k^iB|| < \infty$ if $|\lambda| < |\mu_k|$. Hence, we can find a sequence λ_k, such that $\lambda_kA_k \to A_\infty$ and

$$k < \sum_{i=0}^{\infty} |\lambda_k^i|\ ||CA_k^iB|| < \infty.$$

The set, consisting of A_∞ and all $\lambda_k A_k$, is compact, all (C, A, B) are BIBO stable, but $\sum_{i=0}^{\infty} ||CA^iB||$ is not bounded on this set. As is clear from the proof of proposition 4.4, however, it is necessary for BIBO stability of the sequence that there exists a number G, such that

$$\sum_{i=0}^{\infty} ||CA^iB|| < G \quad \forall (C, A, B) \in D.$$

□

4.3 A spectral density for a time-varying AR process

In this section, we introduce a spectral density for a sequence of time varying AR processes. Dahlhaus already did this, but we shall show that by using the concept of BIBO stability for a sequence of time-varying systems, matters get clearer and easier. In fact, we shall show that having a spectral density is the same as being stable.

Consider the following sequence of time varying Gaussian AR(n) processes with increasing number of path points:

$$y_{t,T} = (a_{t,T})_1 y_{t-1,T} + \cdots + (a_{t,T})_n y_{t-n,T} + \sigma_{t,T} e_{t,T} \qquad (0 \leq t \leq T; T \in \mathbb{N}) \tag{4.20}$$

where

- $(e_{0,T}, \cdots, e_{T,T})' \stackrel{d}{=} \mathcal{N}(0, I_{T+1})$;
- $(y_{0,T}, y_{-1,T}, \cdots, y_{-n,T})' \stackrel{d}{=} \mathcal{N}(0, V_T)$ where V_T is positive definite and Toeplitz, such that

$$V_T \begin{pmatrix} 1 \\ (a_{0,T})_1 \\ \vdots \\ (a_{0,T})_n \end{pmatrix} = \begin{pmatrix} \sigma_{0,T}^2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}; \tag{4.21}$$

- $(y_{-1,T}, \cdots, y_{-n,T})$ independent of $(e_{0,T}, \cdots, e_{T,T})$ for all $T \in \mathbb{N}$;
- $\sigma_{t,T} > 0$ for all t, $(0 \leq t \leq T)$ and all $T \in \mathbb{N}$. So, we can write: $\sigma_{t,T} = e^{r_{t,T}}$.

Note that because of (4.21) the polynomial with natural coordinates $(a_{0,T}, \sigma_{0,T}^2)$ is an element of U_n, i.e. it is stable, it has degree n, and its highest coefficient is positive.
The processes $y_{\cdot,T}$ can be extended to be defined for all $t \in \mathbb{Z}; t \leq T$, such that the processes remain Gaussian and $\boldsymbol{E} y_{t_1-s,T} y_{t_1,T} = \boldsymbol{E} y_{t_2-s,T} y_{t_2,T}$ for all $s \in \mathbb{Z}_+; t_1, t_2 \leq 0$. In fact, the processes in the time before 0, $(y_{t,T})_{t \leq 0}$, are just constant AR(n) processes with coefficients and noise levels $(a_{0,T}, \sigma_{0,T}^2)$. In the smallest Hilbert space H_{y_T}, which

contains all $y_{t,T}$ $(t \leq T)$, we can extend the set $\{e_{t,T} | 0 \leq t \leq T\}$ to an orthonormal basis $(e_{t,T})_{t \leq T}$. We can define an isometry on this Hilbert space into $L_2(sT)$ by:

$$e_{t,T} \mapsto \begin{pmatrix} \mathbf{T} & \rightarrow & \mathcal{C} \\ x & \mapsto & \bar{x}^t \end{pmatrix}. \tag{4.22}$$

Under this isometry, we have the correspondence:

$$y_{t,T} \mapsto \bar{x}^t p_{t,T}(x) \tag{4.23}$$

where $z \mapsto p_{t,T}(z)$ is a Hardy function. The isometric property of this map can be formulated as:

$$\boldsymbol{E} y_{t,T} y_{s,T} = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i(s-t)\phi} p_{t,T}(e^{i\phi}) \overline{p_{s,T}(e^{i\phi})} d\phi = \int_{\mathbf{T}} \bar{x}^{t-s} p_{t,T}(x) \overline{p_{s,T}(x)} dm(x), \tag{4.24}$$

and the map translates equation (4.20) in:

$$p_{t,T}(x) = (a_{t,T})_1 x p_{t-1,T}(x) + \cdots + (a_{t,T})_n x^n p_{t-n,T}(x) + \sigma_{t,T}. \tag{4.25}$$

We shall make two assumptions about the coefficients and noise levels. First, a definition.

Definition. We shall call a finite sequence of vectors in $\mathbf{R}^p$ a *path* in $\mathbf{R}^p$. Let be given a sequence of paths $(b_{t,T})_{0 \leq t \leq T,\ T \in \mathbf{N}}$ in $\mathbf{R}^p$. The second index T indicates the number of points of the path $(b_{t,T})_{0 \leq t \leq T}$. We state that the $b_{t,T}$ vary *equislowly* with increasing number of path points, if there is a number C, such that

$$||b_{t+1,T} - b_{t,T}|| < \frac{C}{T} \text{ for all } t, \quad 0 \leq t \leq T-1. \square$$

Assumption 1.

> The initial log noise levels $r_{0,T} = \log(\sigma_{0,T})$ are uniformly bounded. The coefficients $a_{t,T}$, and log noise levels $r_{t,T} = \log(\sigma_{t,T})$ vary equislowly with the increasing number of path points.

Because of (4.21), the initial coefficients $a_{0,T}$ are in AR_n, hence uniformly bounded. Uniform boundedness of the initial log noise levels $r_{0,T}$ and the initial coefficients $a_{0,T}$, together with the equislowness-property, ensure that all coefficients $a_{t,T}$ and log noise levels $r_{t,T}$ $(0 \leq t \leq T)$ are uniformly bounded. The log noise levels are uniformly bounded if and only if the noise levels are uniformly bounded away from zero and infinity.
We could have replaced the ordinary metric $||\ ||$ in Assumption 1 by any other metric $\boldsymbol{d}_n$ on a compact set $K \subset \mathbf{R}^{n+1}$, containing all coefficients $\begin{pmatrix} a_{t,T} \\ r_{t,T} \end{pmatrix}$, without altering the assumption. This, provided it satisfies the following property: there exist positive numbers ε_K and M_K, such that

$$\boldsymbol{d}(p,q) < M_K ||p-q|| \quad \text{if} \quad ||p-q|| < \varepsilon_K \tag{4.26}$$

$$||p-q|| < M_K \boldsymbol{d}(p,q) \quad \text{if} \quad \boldsymbol{d}(p,q) < \varepsilon_K \tag{4.27}$$

for all $p, q \in K$ [4] Consequently, if the noise levels are uniformly bounded away from zero and infinity, then the log noise levels $r_{t,T}$ satisfy the equislowness property, if and only if the noise levels $\sigma^2_{t,T}$ (or their roots $\sigma_{t,T}$) have this property.

Let the polygons $\gamma_T = \begin{pmatrix} a_T \\ \sigma^2_T \end{pmatrix}$ be defined on $(-\infty, 1]$ as follows:

$$\begin{pmatrix} a_T(u) \\ \sigma^2_T(u) \end{pmatrix} = \begin{pmatrix} a_{0,T} \\ \sigma^2_{0,T} \end{pmatrix} \text{ if } u \leq 0; \tag{4.28}$$

$$\begin{pmatrix} a_T(u) \\ \sigma^2_T(u) \end{pmatrix} = \begin{pmatrix} a_{t,T} \\ \sigma^2_{t,T} \end{pmatrix} + (Tu - t) \times (\begin{pmatrix} a_{t+1,T} \\ \sigma^2_{t+1,T} \end{pmatrix} - \begin{pmatrix} a_{t,T} \\ \sigma^2_{t,T} \end{pmatrix}) \tag{4.29}$$

$$\text{if, } \frac{t}{T} < u \leq \frac{t+1}{T}, 0 \leq t \leq T.$$

These polygons are constant left of 0, equicontinuous (even equiLipschitz), and $(\begin{pmatrix} a_T(u) \\ \sigma^2_T(u) \end{pmatrix})_{T \in \mathbf{N}}$ is bounded for every $u \in (-\infty, 1]$. Hence, the Theorem of Arzela Ascoli applies: every subsequence of these polygons has a subsubsequence that uniformly converges to a continuous curve on $(-\infty, 1]$.

Assumption 2.

For all $u \in [0,1]$ $\lim_{T \to \infty} \begin{pmatrix} a_{[uT],T} \\ \sigma^2_{[uT],T} \end{pmatrix}$ exists.

According to Assumption 1, this means that there is only one limit curve $\gamma = \begin{pmatrix} a \\ \sigma^2 \end{pmatrix}$ for the set of polygons, and we have:

$$\lim_{T \to \infty} \begin{pmatrix} a_T(u) \\ \sigma^2_T(u) \end{pmatrix} = \begin{pmatrix} a(u) \\ \sigma^2(u) \end{pmatrix}, \text{ uniformly in } u \in (-\infty, 1].$$

The curve γ is continuous and of bounded variation. In fact, it is easy to see that if the polygons γ_T converge to γ uniformly in u, then γ is of bounded variation if and only if the coefficients and log noise levels vary equislowly.

We study the following form for a spectral density, introduced by Dahlhaus[5]:

$$f(u,w) = \lim_{T \to \infty} \sum_{s=-\infty}^{\infty} w^s \boldsymbol{Cov}(y_{[uT+\frac{s}{2}],T}, y_{[uT-\frac{s}{2}],T}), \tag{4.30}$$

with $u \in (-\infty, 1]$, $w \in \mathbf{T}$, i.e. $w = e^{i\psi}$, $\psi \in [-\pi, \pi]$ and the summation is taken over all s, where $\boldsymbol{Cov}(y_{[uT+\frac{s}{2}],T}, y_{[uT-\frac{s}{2}],T})$ is defined.

[4] Such a property is certainly satisfied by any Riemannian metric defined on a differential manifold M, which has a parametrization defined on an open set U covering K, see chapter two.

[5] In his introduction of this form, Dahlhaus refers to Wigner-Ville

Due to our two assumptions, we are able to give the exact condition under which this spectral density exists and is non-negative everywhere.
Technical observation: in the proposition 4.6 and the proof below, we use the following notational convention for the unit circle variables w and x:

$$w = e^{i\psi}, \quad (\psi \in [-\pi, \pi]); \quad x = e^{i\phi}, \quad (\phi \in [-\pi, \pi]).$$

Proposition 4.6 The sequence of time-varying AR(n) processes with increasing number of path points, defined in (4.20) and satisfying Assumptions 1 and 2, admits Dahlhaus' spectral density (4.30) in the sense that

$$0 \leq f(u,w) = \lim_{T\to\infty} \sum_{s=-\infty}^{\infty} w^s \boldsymbol{Cov}(y_{[uT+\frac{r}{2}+\frac{s}{2}],T}, y_{[uT+\frac{r}{2}-\frac{s}{2}],T}),$$
$$\text{uniformly in } w, \ |w| = 1 \text{ for all } r \in \mathbf{Z}, \tag{4.31}$$

if and only if all points of the limit curve γ are stable, i.e. $\gamma(u) \in U_n$ for all $u \in (-\infty, 1]$. Moreover, we have:

$$f(u,w) = \frac{\sigma(u)^2}{|1 - a(u)_1 w + \cdots + -a(u)_n w^n|^2}.$$

Proof:
if.
With the introduction of the polygons γ_T, it is possible to define a type of "polygons" through the transfer functions $p_{t,T}$ by expanding equation (4.25) into:

$$p_T(u,x) - a_T(u)_1 x p_T(u - \frac{1}{T}, x) + \cdots - a_T(u)_n x^n p_T(u - \frac{n}{T}, x) = \sigma_T(u). \tag{4.32}$$

Given $\gamma_T = (a_T, \sigma_t^2)$, we can solve this equation, and we find a unique solution p_T, such that $p_T(u,x) = p_T(v,x)$ for all $u, v \leq 0$:

$$\text{if } u \leq 0 \text{ then } p_T(u,x) = \frac{\sigma(0)}{1 - a(0)_1 x + \cdots + -a(0)_n x^n},$$

$$\text{if } 0 < u \leq \tfrac{1}{T}, \text{ then } p_T(u,x) = a(u)_1 x p_T(0,x) + \cdots + a(u)_n x^n p_T(0,x) + \sigma(u) \text{ etc.}$$

Note that $p_T(u,x)$ is rational and well defined in x for all $u \in (-\infty, 1]$. Hence, $\phi \mapsto p_T(u, e^{i\phi})$ is a many times differentiable function on $(-\pi, \pi)$ for all u. Note also that $p_{t,T}(x) = p_T(\frac{t}{T}, x)$.
We shall prove that $p_T(u, e^{i\phi})$ and $\frac{d^2}{d\phi^2} p_T(u, e^{i\phi})$ are uniformly bounded in u and ϕ, and that the set of functions $(p_T(u,x))_{T\in\mathbf{N}, x\in\mathbf{T}}$ is equiLipschitz in u.
Choose $\varpi \leq 0$. Define:

$$A_{k,T} = \begin{pmatrix} a_T(\varpi + \frac{k+1}{T})_1 x & \cdots & \cdots & a_T(\varpi + \frac{k+1}{T})_n x^n \\ 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \end{pmatrix};$$

$$B = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}; C = (1\ 0\ \cdots 0).$$

Then, the sequence of systems with increasing number of path points is:

$$\left\{ \begin{array}{lcl} X_{k+1,T} & = & A_{k,T}X_{k,T} + Bu_{k,T} \\ z_{k,T} & = & CX_{k,T} \end{array} \right\} \quad (0 \leq k \leq T) \tag{4.33}$$

is BIBO stable, according to proposition 4.4, and the assumption that $\gamma(u)$ is stable for all u. It is also clear from the proof of that proposition and lemma 4.2 that the constants in the definition of BIBO stability don't depend on ϖ nor on x. The same is true for the sequence of systems that we obtain by replacing the map $B : \mathbb{R}^m \to \mathbb{R}^n$ by $A_{0,T} : \mathbb{R}^n \to \mathbb{R}^n$. Hence, bounded initial states in (4.33) yield a bounded contribution to the outputs. Now, define

$$X_{k,T} = \begin{pmatrix} p_T(\varpi + \frac{k}{T}, x) \\ \vdots \\ p_T(\varpi + \frac{k-(n-1)}{T}, x) \end{pmatrix};$$

$$z_{k,T} = p_T(\varpi + \frac{k}{T}, x); \quad u_{k,T} = \sigma_T(\varpi + \frac{k+1}{T}).$$

As $z_{k,T}$ is the output of system (4.33), if $u_{k,T}$ is the input, and this input and the initial states are bounded, also $z_{k,T}$ is uniformly bounded in ϖ, k, T, x. Hence, $p_T(u, x)$ is uniformly bounded in u and x.
Next, we differentiate equation (4.32) with respect to ϕ (remember that we write $x \in \mathbf{T}$ instead of $e^{i\phi}$, $\phi \in [-\pi, \pi]$). We write $p'_T(u, x)$ instead of $\frac{d}{d\phi}p_T(u, e^{i\phi})$. We obtain:

$$p'_T(u,x) - a_T(u)_1 x p'_T(u - \frac{1}{T}, x) + \cdots - a_T(u)_n x^n p'_T(u - \frac{n}{T}, x) =$$

$$= ia_T(u)_1 x p_T(u - \frac{1}{T}, x) + \cdots + ina_T(u)_n x^n p_T(u - \frac{n}{T}, x)$$

Define:

$$X_{k,T} = \begin{pmatrix} p'_T(\varpi + \frac{k}{T}, x) \\ \vdots \\ p'_T(\varpi + \frac{k-(n-1)}{T}, x) \end{pmatrix};$$

$$z_{k,T} = p'_T(\varpi + \frac{k}{T}, x);$$

$$u_{k,T} = ia_T(\varpi + \frac{k+1}{T})_1 x p_T(\varpi + \frac{k}{T}, x) + \cdots + ina_T(\varpi + \frac{k+1}{T})_n x^n p_T(\varpi + \frac{k-(n-1)}{T}, x).$$

Again: As $z_{k,T}$ is the output of system (4.33), corresponding with the input $u_{k,T}$, and as well this input as the initial states are bounded, also $z_{k,T}$ is uniformly bounded in

ϖ, k, T, x. Hence, $p'_T(u,x)$ is uniformly bounded in u and x.
It can be proved in the same way that also

$$p''_T(u,x) = \frac{d^2}{d\phi^2} p_T(u, e^{i\phi})$$

is uniformly bounded in u and x. This fact will be used below.
By subtracting from equation (4.25) the same equation, in which u is replaced by v, we obtain:

$$(p_T(u,x) - p_T(v,x)) - a_T(v)_1 x(p_T(u - \frac{1}{T}, x) - p_T(v - \frac{1}{T}, x)) + \cdots +$$

$$-a_T(v)_n x^n (p_T(u - \frac{n}{T}, x) - p_T(v - \frac{n}{T}, x)) =$$

$$= (\sigma_T(u) - \sigma_T(v)) + (a_T(u)_1 - a_T(v)_1) x p_T(u - \frac{1}{T}, x) + \cdots + (a_T(u)_n - a_T(v)_n) x^n p_T(u - \frac{n}{T}).$$

Choose $\varpi, \tilde{\varpi} \leq 0$. Define:

$$X_{k,T} = \begin{pmatrix} p_T(\tilde{\varpi} + \frac{k}{T}) - p_T(\varpi + \frac{k}{T}, x) \\ \vdots \\ p_T(\tilde{\varpi} + \frac{k-(n-1)}{T}, x) - p_T(\varpi + \frac{k-(n-1)}{T}) \end{pmatrix};$$

$$z_{k,T} = p_T(\tilde{\varpi} + \frac{k}{T}, x) - p_T(\varpi + \frac{k}{T}, x);$$

$$u_{k,T} = \sigma_T(\tilde{\varpi} + \frac{k+1}{T}, x) - \sigma_T(\varpi + \frac{k+1}{T}, x) +$$

$$+ \sum_{j=1}^{n} (a_T(\tilde{\varpi} + \frac{k+1}{T})_j - a_T(\varpi + \frac{k+1}{T})_j) x^j p_T(\tilde{\varpi} + \frac{k-(j-1)}{T}, x).$$

We have seen that the polygons $\gamma_T = \begin{pmatrix} a_T \\ \sigma_T \end{pmatrix}$ are equiLipschitz, and $p_T(u,x)$ is uniformly bounded, hence there exists a number C, such that

$$|u_{k,T}| \leq C||\tilde{\varpi} - \varpi|| \text{ for all } k, T.$$

Furthermore, $X_{0,T} = 0$. As $z_{k,T}$ is the output of system (4.33) with input $u_{k,T}$ and zero initial state, we have that, using the first criterion of BIBO stability, the collection of functions on $(-\infty, 1]$, $u \mapsto p_T(u,x)$, $T \in \mathbb{N}$, $x \in \mathbf{T}$ is equiLipschitz. This collection is also uniformly bounded and constant on $(-\infty, 0]$. Therefore, we can apply the theorem of Arzela Ascoli: every sequence in the collection $(p_{T_k})_{k \in \mathbb{N}}$ has a subsequence $(p_{T_{k_l}})_{l \in \mathbb{N}}$ that converges uniformly in u to a limit function $\tilde{p}$. It follows from equation (4.25) that this limit function should satisfy the equation:

$$\tilde{p}(u,x) - a(u)_1 x \tilde{p}(u,x) + \cdots - a(u)_n x^n \tilde{p}(u,x) = \sigma(u). \tag{4.34}$$

As the polynomial $1-a(u)_1x+\cdots-a(u)_nx^n$ has no roots on the unit circle by assumption, this equation has exactly one solution:

$$\tilde{p}(u,x) = p(u,x) = \frac{\sigma(u)}{1 - a(u)_1x + \cdots - a(u)_nx^n}. \tag{4.35}$$

Consequently, there is only one limit function and the sequence $(p_T)_{T\in\mathbf{N}}$ converges uniformly in u to p. The convergence is pointwise in x, but majorized, because p_T is uniformly bounded in u and x.
These two facts: the convergence of p_T to p, which is uniform in u and majorized in x, and the uniform boundedness of $p_T''(u,x)$ in u and x, are enough to show the convergence of the Dahlhaus expression (4.31) of the spectral density, as we shall do now. Rewriting this expression, using equation (4.24), makes clear that we have to prove that the righthand side of:

$$f(u,w) = \lim_{T\to\infty} \sum_{s=-\infty}^{\infty} w^s \int_{\mathbf{T}} \bar{x}^s p_T(\frac{[uT+\frac{r}{2}+\frac{s}{2}]}{T},x)\overline{p_T(\frac{[uT+\frac{r}{2}-\frac{s}{2}]}{T},x)}dm(x) \tag{4.36}$$

converges uniformly in $u \in [0,1], w \in \mathbf{T}$.
We write $q(s,x)$ instead of $p_T(\frac{[uT+\frac{r}{2}+\frac{s}{2}]}{T},x)$. As $q(s,e^{i\phi})$ is a three times continuously differentiable periodic function of ϕ on $[-\pi,\pi]$, we have:

$$q(s,x) = \sum_{k=-\infty}^{\infty} q_k(s)x^k, \text{ and } q''(s,x) = -\sum_{k=-\infty}^{\infty} k^2 q_k(s)x^k \tag{4.37}$$

uniform in $x \in \mathbf{T}$, where $(q_k(s))_{k\in\mathbf{Z}}$'s are the Fourier coefficients of the function $x \mapsto q(s,x)$. As we have seen, $q''(s,x)$ is uniformly bounded in x and s and T, and consequently, there exists a constant C, such that

$$|q_k(s)| < \frac{C}{k^2} \text{ for all } s,T;\ k \neq 0. \tag{4.38}$$

Also, there exists a constant $\tilde{C}$, such that

$$|q_o(s)| \leq \tilde{C} \text{ for all } s,T. \tag{4.39}$$

We first consider the tails in the sum of expression (4.36). We obtain the following inequalities, using (4.38, 4.39) and Parseval's theorem:

$$|\sum_{s=N}^{\infty} w^s \int_{\mathbf{T}} \bar{x}^s q(s,x)\overline{q(-s,x)}dm(x)| \leq \sum_{s=N}^{\infty} |\int_{\mathbf{T}} \bar{x}^s q(s,x)\overline{q(-s,x)}dm(x)| =$$

$$= \sum_{s=N}^{\infty} |\sum_{k\in\mathbf{Z}} q_k(s)\overline{q_{k-s}(-s)}| \leq \sum_{s=N}^{\infty} |q_0(s)|\ |q_{-s}(-s)| +$$

$$+ \sum_{s=N}^{\infty} |q_s(s)|\ |q_0(-s)| + \sum_{s=N}^{\infty} \sum_{k\in\mathbf{Z};k\neq 0;k\neq s} |q_k(s)|\ |q_{k-s}(-s)| \leq$$

$$\leq 2\sum_{s=N}^{\infty}\tilde{C}\frac{C}{s^2}+C\sum_{s=N}^{\infty}\sum_{k\in\mathbf{Z};k\neq 0;k\neq s}\frac{1}{k^2(k-s)^2}=$$

$$=2\tilde{C}C\sum_{s=N}^{\infty}\frac{1}{s^2}+C\sum_{s=N}^{\infty}\frac{1}{s^2}(\sum_{k\in\mathbf{Z};k\neq 0;k\neq s}(\frac{1}{k}-\frac{1}{k-s})^2)\leq(2\tilde{C}C+2C\frac{\pi^2}{3})\sum_{s=N}^{\infty}\frac{1}{s^2}.$$

This shows that for any $\varepsilon>0$ it is possible to find a number N, not depending on T nor on w, such that the following tail is smaller than ϵ:

$$|\sum_{s=N}^{\infty}w^s\int_{\Gamma}\bar{x}^s q(s,x)\overline{q(-s,x)}dm(x)|<\varepsilon.$$

Next, we only have to consider the middle part of the summation:

$$\lim_{T\to\infty}\sum_{s=-N}^{N}w^s\int_{\Gamma}\bar{x}^s q(s,x)\overline{q(-s,x)}dm(x)=\sum_{s=-N}^{N}w^s\int_{\Gamma}\bar{x}^s|p(u,x)|^2dm(x)$$

due to the majorized convergence of $q(s,x)$ to $p(u,x)$ (Lebesgue). This proves that the convergence is uniform in w, and that

$$f(u,w)=|p(u,w)|^2.$$

only if

It suffices to show that

$$\int_{\Gamma}(1-a(u)_1w-\cdots-a(u)_nw^n)\bar{w}^i f(u,w)dm(w)=0 \text{ for all } i=1,\cdots n \text{ for all } u. \quad (4.40)$$

Then, the polynomial $1-a(u)_1w-\cdots-a(u)_nw^n$ is a multiple of the zero evaluator w.r.t. the Schur inner product on Π_n, defined by $f(u,w)$ and must, therefore, have all its roots outside the unit disk. Hence, the curve γ lies completely in U_n.
First, note that:

$$\sum_{s=-\infty}^{\infty}w^s\int_{\Gamma}\bar{x}^s x^j p_T(\frac{[uT+\frac{s}{2}]-j}{T},x)\overline{\bar{x}^i p_T(\frac{[uT-\frac{s}{2}]-i}{T},x)}dm(x)=$$

$$=\sum_{s=-\infty}^{\infty}w^s\int_{\Gamma}\bar{x}^s x^j p_T(\frac{[(uT-\frac{j}{2}-\frac{i}{2})+\frac{s-j+i}{2}]}{T},x)\overline{\bar{x}^i p_T(\frac{[(uT-\frac{j}{2}-\frac{i}{2})-\frac{s-j+i}{2}]}{T},x)}dm(x)=$$

$$=w^{j-i}\sum_{t=-\infty}^{\infty}w^t\int_{\Gamma}\bar{x}^t p_T(\frac{[(uT-\frac{j}{2}-\frac{i}{2})+\frac{t}{2}]}{T},x)\overline{p_T(\frac{[(uT-\frac{j}{2}-\frac{i}{2})-\frac{t}{2}]}{T},x)}dm(x),$$

which converges to: $w^{j-i}f(u,w)$ uniformly in w by our assumptions. Then, the expression

$$\sum_{s=-\infty}^{\infty}w^s\int_{\Gamma}\bar{x}^s(p_T(\frac{[uT-\frac{s}{2}]}{T},x)+$$

$$-\sum_{j=1}^{n}a_T(\frac{[uT-\frac{s}{2}]}{T})_j x^j p_T(\frac{[uT-\frac{s}{2}]-j}{T},x))\overline{\bar{x}^i p_T(\frac{[uT-\frac{s}{2}]-i}{T},x)} \quad (4.41)$$

converges to

$$(1 - a(u)_1 w + \cdots a(u)_n w^n)\bar{w}^i f(u,w),$$

uniform in w. Now, the expression (4.41) is equal to:

$$\sum_{s=-\infty}^{\infty} w^s \int_{\mathbf{I}} \bar{x}^s \sigma_T(\frac{[uT - \frac{s}{2}]}{T}) \overline{\bar{x}^i p_T(\frac{[uT - \frac{s}{2}] - i}{T}, x)} dm(x) = \overline{\sum_{s=1}^{\infty} t_s w^s}$$

for some $(t_s)_{s \in \mathbf{N}}$ because, as argued in the beginning of this section, the $p_{t,T}$,s are all Hardy functions. Uniform convergence in w now proves equation (4.40). This concludes the proof of proposition 4.6. □

It is easy to establish the "if" part of this proposition also for sequences of time-varying ARMA processes under the same assumptions; the "only if" part looks a little bit harder.

Corollary 4.7 For a sequence of time-varying AR(n) processes with increasing number of path points, defined in (4.20), satisfying Assumptions 1 and 2, and with time-varying spectral density $f(u,w)$ ($u \in [0,1]$; $w \in \mathbf{T}$), the asymptotic autocovariances are given by:

$$\lim_{T \uparrow \infty} \boldsymbol{Cov}(y_{[uT-j],T}, y_{[uT-l],T}) = \int_{\mathbf{I}} \bar{w}^{l-j} f(u,w) dm(w) \quad \text{for all } l,j \in \mathbf{Z}; \text{ for all } u \in [0,1]. \tag{4.42}$$

Proof:
Because summation and convergence are uniform in w, according to proposition (4.6), we have:

$$\begin{aligned} \int_{\mathbf{I}} \bar{w}^k f(u,w) dm(w) &= \lim_{T \uparrow \infty} \int_{\mathbf{I}} \bar{w}^k \sum_{s=-\infty}^{\infty} w^s \boldsymbol{Cov}(y_{[uT+\frac{r}{2}+\frac{s}{2}],T}, y_{[uT+\frac{r}{2}-\frac{s}{2}],T}) dm(w) = \\ &= \lim_{T \uparrow \infty} \boldsymbol{Cov}(y_{[uT+\frac{r}{2}+\frac{k}{2}],T}, y_{[uT+\frac{r}{2}-\frac{k}{2}],T}) \end{aligned}$$

for all $r,k \in \mathbf{Z}$. From this, equation (4.42) is immediate. □

4.4 A sequence of stochastic coefficient processes

In this section, we construct *sequences* of stochastic coefficient processes, modelled according to the General Model of chapter three. We will show that such a sequence of time-varying AR processes admits the time-varying spectral density, conditional on the coefficient process. This result may help us to describe the asymptotic conditional distribution, given the coefficients of these time-varying AR processes.

In the preceding sections we dealt with the theory of BIBO stability and time-varying spectral density for a sequence of *deterministic* coefficient processes under the assumption that the sequence of processes satisfies the equislowness property. This equislowness property guaranteed the existence of limiting continuous coefficient curves having

bounded variation. This was essential in the proofs above.
In the following sections, we consider *stochastic* coefficient processes. Suppose that we try to embed a sequence of coefficient *random walks* in a continuous coefficient process. Then this process would be a Brownian motion. A Brownian motion, however, does *not* have bounded variation. So, the theory is not directly applicable to that situation. However, the situation is different if we work with a sequence of coefficient *smoothly integrated random walks.* If suitably defined, a limiting curve will be an integrated Ornstein Uhlenbeck Velocity process, which is even differentiable. This idea will be pursued in the next sections. It will lead to a model for a sequence of time-varying AR processes with stochastic coefficients that follow smoothly integrated random walks.
In this section (4.4), we introduce the model for the sequence of stochastic coefficient processes. For every single process in the sequence, we earlier defined its *running time.* For the sequence as a whole, we shall introduce the *asymptotic relative running time.* Its properties will be discussed, and the existence of a limit coefficient curve will be proved.
In section (4.5), we consider a sequence of time-varying AR processes corresponding to the coefficient processes of section (4.4). We shall show that such a sequence admits the time-varying spectral density, introduced above, conditional on the values of the asymptotic relative running time. The whole density will be conditional on the Ornstein Uhlenbeck Velocity (OUV) process that generates the coefficients. In section (4.6), we indicate how the time-varying spectral density could be used in identification and for other purposes.
Some aspects about Ornstein-Uhlenbeck Velocity processes can been found in chapter two.

A sequence of coefficient processes according to the General Model. Main Model Sequence.
Let $X_1, \cdots, X_m$ be m C^∞-vector fields on the n-dimensional Riemannian manifold M.
Let β_i, τ_i $(1 \leq i \leq m)$ be some positive numbers.
Furthermore, for all $T \in \mathbb{N}$, let $q_{0,T} : \Omega \to M$ be stochastic variables, such that $P(q_{0,T} \in U) > 0$ for all open sets $U \subset M$.
Let $(\theta_i(u))_{u \in \mathbb{R}}, (1 \leq i \leq m)$ be m (continuous-time!) Ornstein-Uhlenbeck Velocity processes, mutually independent and independent of $q_{0,T}$ for all $T \in \mathbb{N}$, where $(\theta_i(u))_{u \in \mathbb{R}}$ is determined by the parameters β_i and τ_i, i.e. let θ_i be an OUV(β_i, τ_i)-process.

Then, we recursively define the following sequence of polygonal processes:

$$\text{for all } t \in \mathbf{Z}_+ :$$

$$\theta_{t,T} = \begin{pmatrix} \theta_1(\frac{t}{T}) \\ \vdots \\ \theta_m(\frac{t}{T}) \end{pmatrix}, \tag{4.43}$$

$$\mathcal{B}_{t,T} = \text{the smallest } \sigma\text{-algebra, such that } q_{0,T}, \text{ and } \theta_{s,T} \ (s \leq t) \text{ are measurable,} \tag{4.44}$$

$$d_{t,T} = \frac{1}{T} \sum_{i=1}^{m} \theta_{i\ t,T} X_i, \tag{4.45}$$

$$\begin{aligned} q_{t+1,T} &= \exp_{q_{t,T}}(d_{t,T}) \text{ if this is well defined. If not,} \\ q_{t+1,T} &= h_{t+1,T} \text{ for some } \mathcal{B}_{t,T}\text{-measurable } h_{t+1,T}. \end{aligned} \tag{4.46}$$

□

The running time of process T will be denoted by ν_T.

Lemma 4.8 The polygonal process T, $(q_{t,T}, \frac{1}{T}\sum_{i=1}^{m}\theta_{ti,T}X_i(q_{t,T}))_{t\in\mathbf{Z}_+}$ is of independent directional variation until ν_T if and only if

$$q \mapsto \langle X_i(q), X_j(q)\rangle_q \text{ is constant on } M. \tag{4.47}$$

If the polygonal processes of the sequence are of independent directional variation, then they are modelled according to the General Model (see section (3.7))

Proof: Consider process number T. We have seen in chapter two that $(\theta_{t,T})_{t\in\mathbf{Z}}$ is an AR(1) process, satisfying the equation:

$$\begin{aligned} \theta_{t+1,T} = &\begin{pmatrix} e^{-\frac{\beta_1}{T}} & 0 & \cdots & 0 \\ 0 & e^{-\frac{\beta_2}{T}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \\ 0 & 0 & \cdots & e^{-\frac{\beta_m}{T}} \end{pmatrix}\theta_{t,T} + \\ &+\begin{pmatrix} \sqrt{1-e^{-2\frac{\beta_1}{T}}}\tau_1 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & \sqrt{1-e^{-2\frac{\beta_m}{T}}}\tau_m \end{pmatrix}\lambda_{t,T}, \end{aligned} \tag{4.48}$$

where $(\lambda_{t,T})_{t\in\mathbf{Z}_+}$ is m-dimensional Gaussian white noise with variance equal to the identity, independent of $q_{0,T}$ and $\theta_{0,T}$. The conditions of the proposition (3.4) (integrability, never-evasiveness, proper representability of the variation term) are easily established, hence we can apply this proposition, proving the first statement. Consequently, if we want to prove the second statement, we may assume that the vector fields X_i $(1 \le i \le m)$ have constant inner products on the manifold M. By means of a Gramm-Schmidt procedure, we construct an orthonormal set of vector fields E_j $(1 \le j \le r)$, such that $X_i = C_{ji}E_j$ for some $r \times m$-dimensional matrix C, which is constant on the manifold M. Let Q be the conditional variance of $\theta_{t,T}$, given the past, i.e. let Q be the m-dimensional diagonal matrix with the numbers $\frac{(1-e^{-2\frac{\beta_i}{T}})\tau_i^2}{T^2}$ as entries on its main diagonal. Let U and W be r- and m-dimensional orthogonal matrices, respectively, such that

$$UC\sqrt{Q}W = \begin{pmatrix} \tilde{\tau}_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \tilde{\tau}_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \tilde{\tau}_r & 0 & \cdots & 0 \end{pmatrix}, \text{ where } \tilde{\tau}_i > 0.$$

Let $F_k = \sum_{j=1}^{r} U_{kj}E_j$. Then, $F_1, \cdots, F_r$ form an orthonormal set of r vector fields. Let $\tilde{\theta}_t = W^*\sqrt{Q}^{-1}\theta_{t,T}$. Consequently, $(\tilde{\theta}_t)_{t\in\mathbf{Z}}$ is an m-dimensional Gaussian stationary

process with unit conditional variance given the past, and with an everywhere positive definite spectral density matrix. Now,

$$d_t = \frac{1}{T}\sum_{i=1}^{m}\theta_{ti,T}X_i = \frac{1}{T}\sum_{i=1}^{m}\sum_{j=1}^{r}\theta_{ti,T}C_{ji}E_j = \frac{1}{T}\sum_{i=1}^{m}\sum_{j,k=1}^{r}(\sqrt{Q}W\tilde{\theta}_t)_i C_{ji}U_{kj}F_k =$$

$$= \sum_{k=1}^{r}(UC\sqrt{Q}W\tilde{\theta}_t)_k F_k = \sum_{k=1}^{r}\tilde{\theta}_{tk}\tilde{\tau}_k F_k,$$

and we obtain expression (3.37) in the definition of the General Model. □

From now on we assume that the processes in the sequence are of independent directional variation until their running time ν_T, or, equivalently, that condition (4.47) is satisfied.

Remark 4.9 Consider the role of the independent variation term in changing $q_{t,T}$ into $q_{t+1,T}$ in comparison with the total changing factor:

$$\frac{\boldsymbol{E}||\frac{1}{T}\sum_{i=1}^{m}\sqrt{1-e^{-2\frac{\beta_i}{T}}}\tau_i\lambda_{ti,T}X_i||^2}{\boldsymbol{E}||\frac{1}{T}\sum_{i=1}^{m}\theta_{ti,T}X_i||^2} = \frac{\frac{1}{T^2}\sum_{i=1}^{m}||X_i||^2\tau_i^2(1-e^{-2\frac{\beta_i}{T}})}{\frac{1}{T^2}\sum_{i=1}^{m}||X_i||^2\tau_i^2}.$$

This proportion tends to zero if T tends to infinity. Interpretation: in this sequence of stochastic coefficient processes, the coefficients get ever more "deterministic". □

By introducing the OUV-process θ, we are relating all the elements of the sequence of polygonal processes with each other. The higher the index T of the polygonal process, the more values of the OUV-process θ in the interval $[0,1]$ it uses for the generation of the process points $(q_t, d_t)_{0\le t\le T}$. Following the idea of Dahlhaus, the higher the index T of the polygonal process, the more information of the same kind the polygonal process contains. On account of this, it must be possible to do *asymptotics* on this sequence of polygonal coefficient processes and the corresponding sequence of time-varying AR processes.

Next, assume that the sequence of initial coefficients $q_{0,T}$ of all processes converges, i.e.

$$\lim_{T\uparrow\infty} q_{0,T} = q_0 \text{ with probability 1.} \tag{4.49}$$

Then $q_0 : \Omega \to M$ is $\mathcal{A}$ -measurable.
Now, consider the following increasing sequence of σ-algebras $(\mathcal{B}_u)_{u\in\mathbb{R}_+}$, defined by:

> if $u \ge 0$, then $\mathcal{B}_u$ is the smallest σ-algebra in $\mathcal{A}$ with respect to which q_0; $q_{0,T}$ $(T \in \mathbb{N})$, and $\theta_i(s)$ $(1 \le i \le m,\ 0 \le s \le u)$ are measurable.

Obviously, the previously defined σ-algebras $\mathcal{B}_{t,T}$ satisfy $\mathcal{B}_{t,T} \subset \mathcal{B}_{\frac{t}{T}}$.
Furthermore, we call μ a *stopping time* w.r.t. $(\mathcal{B}_u)_{u\in\mathbb{R}_+}$, if $\mu : \Omega \to \mathbb{R}_+$ is measurable with respect to $\mathcal{A}$, and the set $\{\omega \in \Omega |\ \mu(\omega) > u\}$ is element of $\mathcal{B}_u$ for all $u \in \mathbb{R}_+$. For example, every running time ν_T induces the stopping time $\frac{\nu_T}{T}$ w.r.t. $(\mathcal{B}_u)_{u\in\mathbb{R}_+}$.

The supremum of a sequence of stopping times is again a stopping time. The minimum of a finite sequence of stopping times is again a stopping time. Any *constant* variable: $\Omega \to \mathbb{R}_+$ is a stopping time.

Definition of limit coefficient curve γ and asymptotic relative running time ν. Consider the differential equation on the manifold M [6]:

$$\dot{\gamma}(u) = \sum_{i=1}^{m} \theta_i(u,\omega) X_i(\gamma(u)); \qquad \gamma(0) = q_0. \tag{4.50}$$

The sample path $t \mapsto (\theta_1(t,\omega), \cdots, \theta_m(t,\omega))$ is continuous with probability 1 (see chapter two). Consequently, the right hand side of this differential equation is then also continuous and locally Lipschitz in the argument of the X_i (use charts). This is enough to guarantee the existence and uniqueness of a solution of (4.50). Hence, there is P-a.s. a unique maximal solution γ of (4.50). Let $I(\omega) = (-a(\omega), \nu(\omega))$ be its domain. By definition, γ is C^1 with probability 1 on $I(\omega)$. In the following proposition 4.10, we shall prove that

$$\lim_{T\to\infty} q_{[uT],T} = \gamma(u) \text{ for all } u \in [0, \nu(\omega)), \text{ P-a.s..} \tag{4.51}$$

Hence, γ will be the *limit coefficient curve* of the sequence of processes.
We shall also prove that ν, the supremum of the domain of γ, is a stopping time w.r.t. $(\mathcal{B}_u)_{u\in\mathbb{R}_+}$. We shall call ν the *asymptotic relative running time* of the sequence of polygonal processes. The term can be explained by the fact, as we shall prove, that the running times ν_T of process number T asymptotically satisfy $\nu_T \geq \nu T \ \ (T \to \infty)$. □

In order to prepare the proof of the proposition, we introduce a kind of *polygons* through the points $q_{t,T}$ as follows:
Define the curves of stochastic variables $\gamma_T(u)$ on M by:

$$\gamma_T : [0, \frac{\nu_T}{T}] \times \Omega \ \to \ M \tag{4.52}$$

$$\gamma_T(0,\omega) \ = \ q_{0,T}(\omega) \tag{4.53}$$

$$\gamma_T(u,\omega) \ = \ \exp_{q_{t,T}(\omega)}((u - \frac{t}{T}) \sum_{i=1}^{m} \theta_i(\frac{t}{T},\omega) X_i(q_{t,T})) \tag{4.54}$$

$$\text{if } \frac{t}{T} < u \leq \frac{t+1}{T}$$

Note that $\gamma_T(u)$ is measurable w.r.t. $\mathcal{B}_{\frac{[uT]}{T}}$, hence also w.r.t. $\mathcal{B}_u$.

Proposition 4.10 is in fact a version of the existence theorem of Cauchy and Peano for solutions of an ordinary differential equation. Before we formulate and prove it, first an observation that will be frequently used:

$$||\sum_{i=0}^{m} \theta_i(u,\omega) X_i(q)||_q \text{ is constant on the manifold M}$$

[6] Here, we have written the argument ω in order to stress that $\theta_i(u)$ is a measurable map $\theta_i(u) : \Omega \to M$. We shall frequently omit this argument, so $q_{t,T}$ is used to abreviate $q_{t,T}(\omega)$, and $\gamma(u)$ to abreviate $\gamma(u,\omega)$.

for all $u \in \mathbb{R}_+, \omega \in \Omega$. Consequently, we may omit the dependency of the tangent space in which the norm is taken and just write:

$$||\sum_{i=0}^{m} \theta_i(u,\omega)X_i|| \text{ or merely } ||\sum_{i=0}^{m} \theta_i(u)X_i||.$$

Furthermore, we use the notation $f \wedge g$ for the minimum of two functions f and g, hence

$$(f \wedge g)(x) := \min(f(x), g(x));$$

$$(f \wedge g \wedge h)(x) := \min(f(x), g(x), h(x)).$$

By $\boldsymbol{d}(p,q)$ we mean the Riemannian distance between points p and q on the manifold M. By $\bar{B}_r(p)$ we mean the ball in M with center p and radius r:

$$\bar{B}_r(p) = \{q \in M | \boldsymbol{d}(p,q) \leq r\}.$$

Proposition 4.10 Let γ_T's be the "polygons", as defined above, and ν_T's the running times. Let γ be the maximal solution of the differential equation (4.50) and ν the supremum of its domain.
Suppose that

$$\lim_{T\uparrow\infty} q_{0,T} = q_0 \tag{4.55}$$

with probability 1, *and* in mean square, i.e.

$$\lim_{T\uparrow\infty} \boldsymbol{E}\boldsymbol{d}^2(q_{0,T}, q_0) = 0. \tag{4.56}$$

Then, we have:

1. With probability 1,
$$\liminf_{T\uparrow\infty} \frac{\nu_T}{T} \geq \nu.$$

2. ν is a stopping time w.r.t. $(\mathcal{B}_u)_{u\in\mathbb{R}_+}$. If M is complete, then with probability 1, $\nu = \infty$; $\nu_T = \infty$ for all $T \in \mathbb{N}$. If $\nu < \infty$, then there is a unique way to define γ as a continuous curve on the segment $[0,\nu]$, where $\gamma(\nu)$ lies on the boundary of M in the metric completion of M.

3. Let $R \in \mathbb{R}_+$. Let $\mu = R \wedge \nu$ and $\mu_T = R \wedge \nu \wedge \frac{\nu_T}{T}$. Then μ and μ_T are stopping times, and we have:
$$\lim_{T\uparrow\infty} \gamma_T(u \wedge \mu_T) = \gamma(u \wedge \mu) \text{ uniform in } u \in [0,R] \tag{4.57}$$
with probability 1 *and* in mean square.

Remark 4.11 The number R (*"relative consideration time"*) indicates that we only want to consider the first $[RT]$ elements of the process with index T in the sequence. In fact, the case $R = 1$ has been considered in the previous sections, where the index T of the process was the same as the number of elements considered. By re-indexing, we can always return to the case $R = 1$. □

Proof of the proposition: Since the proof is rather tedious it will be split up into a number of steps.

Step 1. There exists a positive number $s(\omega)$, such that

> With probability 1, the "polygons" γ_T are defined on $[0, s(\omega)]$ for all $T > N$ for some number $N \in \mathbb{N}$ and $\lim_{T\uparrow\infty} \gamma_T(u) = \gamma(u)$ uniform in u, $u \in [0, s(\omega)]$.

(4.58)

Necessarily, the number $s(\omega)$ is strictly smaller than $\nu(\omega)$, $(s(\omega) < \nu(\omega))$.
Proof of Step 1: For every point $p \in M$, there exists an open set $U(p)$, containing the compact ball $\bar{B}_{r_p}(p)$, such that $U(p)$ is a normal neighbourhood of each of its points.[7] This means that there is an open set V in M_q for any $q \in U(p)$, such that $\exp_q : V \to U(p)$ is a diffeomorfism. Particularly, if $d(p, q) < r < r_p$, then $\exp_q$ is still defined on the ball with center 0 and radius $r_p - r$ in the tangent space M_q.
Let $U(q_0(\omega))$ and $\bar{B}_{r_{q_0(\omega)}}(q_0(\omega)$ be such an open set and compact ball around $q_0(\omega)$, respectively. Choose $N_0(\omega)$, such that for all $T > N_0$, $\gamma_T(0) = q_{0,T} \in \bar{B}_{\frac{1}{2}r_{q_0}}(q_0)$.

- *We construct an interval* $[0, s(\omega)]$, *such that* $\gamma_T(u) \in \bar{B}_{r_{q_0}}(q_0)$ *for all* $u \in [0, s(\omega)]$. *Consequently, the set* $\{\gamma_T(u) | T \in \mathbb{N}\}$ *will have* compact closure *in* M *for all* $u \in [0, s(\omega)]$.
  Between two consecutive points $\gamma_T(\frac{t}{T}) = q_{t,T}$ and $\gamma_T(\frac{t+1}{T}) = q_{t+1,T}$ the curves γ_T are geodesics, hence , if $\frac{t}{T} < u < \frac{t+1}{T}$, the length of the tangent vectors is constant:

$$||\dot{\gamma}_T(u)||_{\gamma_T(u)} = ||\sum_{i=1}^{m} \theta_i(\frac{t}{T}) X_i(q_{t,T})||_{q_{t,T}} = ||\sum_{i=0}^{m} \theta_i(\frac{t}{T}) X_i|| \tag{4.59}$$

  (Mind the observation made just before the proposition).
  Define the constant K on the manifold

$$K = \max_{i=1,\cdots,m} ||X_i||. \tag{4.60}$$

  Then, we have the following inequality for the tangents of the polygons:

$$||\dot{\gamma}_T(u)||_{\gamma_T(u)} \leq \sum_{i=1}^{m} |\theta_i(\frac{t}{T})| \max_{i=1,\ldots,m} ||X_i|| = K \sum_{i=1}^{m} |\theta_i(\frac{t}{T})|.$$

[7] In fact a weaker statement than Helgason, I.9.9. or the normal property, vide chapter two.

From this, it follows that for all $0 \leq u \leq \frac{\nu_T}{T}$,

$$d(\gamma_T(u), \gamma_T(0)) \leq \int_0^u ||\dot{\gamma}_T(w)||_{\gamma_T(w)} dw \leq K \frac{1}{T} \sum_{t=0}^{[uT]+1} (\sum_{i=1}^m |\theta_i(\frac{t}{T})|).$$

The functions θ_i are continuous with probability 1, hence integrable. With probability 1 $(\frac{1}{T}\sum_{t=0}^{[uT]+1}(\sum_{i=1}^m |\theta_i(\frac{t}{T})|))_{T \in \mathbf{N}}$ converges to $\int_0^u \sum_{i=1}^m |\theta_i(\tau)| d\tau$ uniform in u, $u \in [0, \frac{\nu_T}{T}]$. The latter is a continuous function of u. For that reason, (with probability 1), there exists a number $s(\omega) > 1$, such that

$$\int_0^u \sum_{i=1}^m |\theta_i(\tau)| d\tau < \frac{r_{q_0}}{2K} \text{ if } 0 \leq u \leq s(\omega), \tag{4.61}$$

and there is a number $N(\omega) > N_0(\omega)$, such that for all $T > N$ and $u \in [0, s(\omega)]$ we have:

$$\frac{1}{T} \sum_{t=0}^{[uT]+1} (\sum_{i=1}^m |\theta_i(\frac{t}{T})|) < \frac{r_{q_0}}{2K}.$$

It is clear now that on the whole intervall $[0, s(\omega)]$ the polygons γ_T's lie in the compact ball $B_{r_{q_0}}(q_0)$.

- *We prove that the set of curves* $V = \{\gamma_T : [0, s(\omega)] \to M \mid T > N(\omega)\}$ *is equicontinuous (even equiLipschitz)*
  Just note that because all $u \mapsto \theta_i(u, \omega)$'s are continuous on $[0, s(\omega)]$, also $\sum_{i=1}^m |\theta_i(u)|$ is maximized by a constant $C(\omega)$ on this interval. We have:

$$||\dot{\gamma}_T(u)||_{\gamma_T(u)} \leq C(\omega)K =: D(\omega) \text{ for all } u \in [0, s(\omega)]., \tag{4.62}$$

and also for all $0 \leq v \leq u \leq s(\omega)$,

$$d(\gamma_T(u), \gamma_T(v)) \leq \int_v^u ||\dot{\gamma}_T(w)||_{\gamma_T(w)} dw \leq D(\omega)|u - v|. \tag{4.63}$$

- Now we can apply the theorem of Arzela-Ascoli (see chapter two). Every sequence in V has a subsequence $(\gamma_{T_k})_{k \in \mathbf{N}}$, which converges to a curve $\tilde{\gamma}$, uniformly on $[0, s(\omega)]$ and with probability 1. *We shall show that every limit curve must satisfy the differential equation (4.50), so they all coincide with* γ. Then, it can be concluded that $\lim_{T \uparrow \infty} \gamma_T(u) = \gamma(u)$ with probability 1, and is uniform in $u \in [0, s(\omega)]$.

- **But first** *we show that, a.s. if* $\frac{t}{T} < u < \frac{t+1}{T}$,

$$\lim_{T \uparrow \infty} ||\dot{\gamma}_T(u) - \sum_{i=1}^m \theta_i(u) X_i(\gamma_T(u))||_{\gamma_T(u)} = 0. \tag{4.64}$$

In order to do this, we derive an inequality for the form $\langle \dot{\gamma}_T(u), \sum_{i=1}^m \theta_i(\frac{t}{T}) X_i(\gamma_T(u)) \rangle_{\gamma_T(u)}$, and after that, an inequality for the form $||\dot{\gamma}_T(u) - \sum_{i=1}^m \theta_i(u) X_i(\gamma_T(u))||^2_{\gamma_T(u)}$.

Note that

$$\lim_{u \downarrow \frac{t}{T}} \langle \dot{\gamma}_T(u), \sum_{i=1}^{m} \theta_i(\frac{t}{T}) X_i(\gamma_T(u)) \rangle_{\gamma_T(u)} =$$

$$= ||\sum_{i=1}^{m} \theta_i(\frac{t}{T}) X_i(q_{t,T})||^2_{q_{t,T}} = ||\sum_{i=1}^{m} \theta_i(\frac{t}{T}) X_i||^2. \tag{4.65}$$

As γ_T is geodesical between subsequent points $q_{t,T}$, and $q_{t+1,T}$, we have that $\nabla_{\dot{\gamma}_T(u)} \dot{\gamma}_T(u) = 0$, hence

$$\frac{\partial}{\partial u}(\langle \dot{\gamma}_T(u), \sum_{i=1}^{m} \theta_i(\frac{t}{T}) X_i(\gamma_T(u)) \rangle_{\gamma_T(u)}) =$$

$$= \langle \nabla_{\dot{\gamma}_T(u)} \dot{\gamma}_T(u), \sum_{i=1}^{m} \theta_i(\frac{t}{T}) X_i(\gamma_T(u)) \rangle_{\gamma_T(u)} +$$

$$+ \langle \dot{\gamma}_T(u), \nabla_{\dot{\gamma}_T(u)} \sum_{i=1}^{m} \theta_i(\frac{t}{T}) X_i(\gamma_T(u)) \rangle_{\gamma_T(u)} =$$

$$= \langle \dot{\gamma}_T(u), \nabla_{\dot{\gamma}_T(u)} \sum_{i=1}^{m} \theta_i(\frac{t}{T}) X_i(\gamma_T(u)) \rangle_{\gamma_T(u)}. \tag{4.66}$$

On every tangent space, we define the symmetrical bilinear form:

$$\left\{ \begin{array}{lll} M_q \times M_q & \rightarrow & \mathbf{R} \\ (a,b) & \mapsto & \langle \nabla_a X_i, b \rangle_q + \langle a, \nabla_b X_i \rangle_q = \langle a, F_i(q) b \rangle_q \end{array} \right\}.$$

These bilinear forms are uniformly bounded on the compact set $\bar{B}_{r_{q_0}}(q_0)$, containing all images $\gamma_T(u)$, because

$$||F_i(q)||_{q,\text{operatornorm}} \leq F(\omega) \quad \forall q \in \bar{B}_{r_{q_0(\omega)}}(q_0(\omega)).$$

Using this expression (and also equations (4.66) and (4.59)) and the boundedness of the θ_i, we conclude that there is a number $G(\omega)$, such that

$$|\frac{\partial}{\partial u} \langle \dot{\gamma}_T(u), \sum_{i=1}^{m} \theta_i(\frac{t}{T}) X_i(\gamma_T(u)) \rangle_{\gamma_T(u)}| =$$

$$= |\langle \dot{\gamma}_T(u), \nabla_{\dot{\gamma}_T(u)} \sum_{i=1}^{m} \theta_i(\frac{t}{T}) X_i(\gamma_T(u)) \rangle_{\gamma_T(u)}| =$$

$$= |\sum_{i=1}^{m} \theta_i(\frac{t}{T}) \langle \dot{\gamma}_T(u), \nabla_{\dot{\gamma}_T(u)} X_i(\gamma_T(u)) \rangle_{\gamma_T(u)}| \leq$$

$$\leq mC(\omega) F(\omega) ||\dot{\gamma}_T(u)||^2 = G(\omega) ||\sum_{i=1}^{m} \theta_i(\frac{t}{T}) X_i||^2$$

if $\frac{t}{T} < u < \frac{t+1}{T}$. This, together with equation (4.65), enables us to derive the following inequality for $\langle \dot{\gamma}_T(u), \sum_{i=1}^{m} \theta_i(\frac{t}{T}) X_i(\gamma_T(u)) \rangle_{\gamma_T(u)}$:

$$\langle \dot{\gamma}_T(u), \sum_{i=1}^{m} \theta_i(\frac{t}{T}) X_i(\gamma_T(u)) \rangle_{\gamma_T(u)} \geq ||\sum_{i=1}^{m} \theta_i(\frac{t}{T}) X_i||^2 -$$

$$-G(\omega)|u-\frac{t}{T}|\|\sum_{i=1}^{m}\theta_i(\frac{t}{T})X_i\|^2.$$

Now, we are able to establish the main inequality. In the derivation, we also use equation (4.62)

$$\|\dot{\gamma_T}(u)-\sum_{i=1}^{m}\theta_i(u)X_i(\gamma_T(u))\|^2_{\gamma_T(u)}=$$

$$=\|\sum_{i=1}^{m}\theta_i(\frac{t}{T})X_i\|^2+\|\sum_{i=1}^{m}\theta_i(u)X_i\|^2-2\langle\dot{\gamma_T}(u),\sum_{i=1}^{m}\theta_i(\frac{t}{T})X_i(\gamma_T(u))\rangle_{\gamma_T(u)}+$$

$$+2\langle\dot{\gamma_T}(u),\sum_{i=1}^{m}(\theta_i(\frac{t}{T})-\theta_i(u))X_i(\gamma_T(u))\rangle_{\gamma_T(u)}\leq$$

$$\leq\|\sum_{i=1}^{m}\theta_i(u)X_i\|^2-\|\sum_{i=1}^{m}\theta_i(\frac{t}{T})X_i\|^2+2D(\omega)\|\sum_{i=1}^{m}(\theta_i(\frac{t}{T})-\theta_i(u))X_i\|+$$

$$+2G(\omega)|u-\frac{t}{T}|\,\|\sum_{i=1}^{m}\theta_i(\frac{t}{T})X_i\|^2\leq$$

$$\leq 4D(\omega)\|\sum_{i=1}^{m}(\theta_i(u)-\theta_i(\frac{t}{T}))X_i\|+2D(\omega)^2G(\omega)|u-\frac{t}{T}|$$

for all $u, \frac{t}{T}<u\leq\frac{t+1}{T}$. The functions $u\mapsto\theta_i(u,\omega)$ are continuous on $[0,s(\omega)]$, so also uniformly continuous. With the inequality just derived, this proves (4.64).

- **Secondly,** *we prove that any limit curve $\tilde{\gamma}$ satisfies the differential equation (4.50).* Let $f: M\to\mathbb{R}$ be a C^∞ function, and let $(\gamma_{T_k})_{k\in\mathbb{N}}$ be a subsequence of the polygons, converging to $\tilde{\gamma}$. Then df, defined by $df_q: M_q\to\mathbb{R}, df_q(v)=v(f)$, is a differential form on M [8]. The matrix elements of df_q are C^∞-functions of q. The matrix norms of the df_q are uniformly bounded on the tangent spaces M_q on any compact subset of M. From (4.64), it follows that

$$\lim_{k\uparrow\infty}df_{\gamma_{T_k}(u)}(\dot{\gamma}_{T_k}(u))-df_{\gamma_{T_k}(u)}(\sum_{i=1}^{m}\theta_i(\frac{[uT]}{T})X_i(\gamma_{T_k}(u)))=0.$$

Hence, also

$$\lim_{k\uparrow\infty}df_{\gamma_{T_k}(u)}(\dot{\gamma}_{T_k}(u))=df_{\tilde{\gamma}(u)}(\sum_{i=1}^{m}\theta_i(u)X_i(\tilde{\gamma}(u))),\text{ uniform in }u\in[0,s(\omega)],$$

or, equivalently,

$$\lim_{k\uparrow\infty}\dot{\gamma}_{T_k}(u)(f)=\sum_{i=1}^{m}\theta_i(u)X_i(\gamma_{T_k}(u))(f),$$

which yields:

$$\lim_{k\uparrow\infty}\int_0^w\dot{\gamma}_{T_k}(u)(f)du=\int_0^w\sum_{i=1}^{m}\theta_i(u)X_i(\tilde{\gamma}(u))(f)du\text{ for all }w\in[0,s(\omega)].$$

[8] See chapter two for differential forms and tangent vectors, considered as derivations.

As

$$\dot{\gamma}_{T_k}(u)(f) = \frac{\partial f \circ \gamma_{T_k}(u)}{\partial u}$$

we also have:

$$\lim_{k\uparrow\infty} \int_0^w \dot{\gamma}_{T_k}(u)(f)du = \lim_{k\uparrow\infty} f \circ \gamma_{T_k}(w) - f(q_0) =$$

$$= f \circ \tilde{\gamma}(w) - f(q_0) = \int_0^w \dot{\tilde{\gamma}}(u)(f)du,$$

and, consequently

$$\dot{\tilde{\gamma}}(u) = \sum_{i=1}^{m} \theta_i(u) X_i(\tilde{\gamma}(u)), \quad \forall u \in [0, s(\omega)] \text{ and } \tilde{\gamma}(0) = q_0$$

i.e., $\tilde{\gamma}$ satisfies the differential equation (4.50).

- Since this differential equation has exactly one solution with initial condition $\gamma(0) = q_0$, so $\tilde{\gamma} = \gamma$. As the maximal solution of equation (4.50) is not defined on $[\nu(\omega), \infty]$, all this also shows that necessarily, $s(\omega) < \nu(\omega)$. This proves Step 1.

Step 2. Statement (4.58) is valid for all intervals $[0, s]$, $0 < s < \nu(\omega)$.

Proof of Step 2: Define the number $S = S(\omega)$, $0 < S(\omega) \leq \infty$ by

$$S(\omega) = \sup\{s \in (0, \nu(\omega))| \text{ statement (4.58) is valid on } [0, s(\omega)]\}.$$

By Step 1. of the proof, it is clear that $S(\omega)$ is well defined (with probability 1). Suppose that $S(\omega) < \nu(\omega)$. Then, the point $\gamma(S)$ exists. We can make a compact ball $\bar{B}_{r_{\gamma(S)}}(\gamma(S))$ around this point, contained in an open set $U(\gamma(S))$, which is a normal neighbourhood of each of its points.
Because the curve γ is continuous, there exists an $\varepsilon > 0$, such that $\gamma(u) \in \bar{B}_{\frac{1}{4}r_{\gamma(S)}}(\gamma(S))$ for all u satisfying $S - \varepsilon < u < S$. We know that the polygons γ_T are equiLipschitz on $[0, S)$ in accordance with (4.63). By definition of S, we know that $\lim_{T\uparrow\infty} \gamma_T(u) = \gamma(u)$ for all $u \in (S - \varepsilon, S)$. Moreover, there exist a δ, $0 < \delta < \varepsilon$ and a number $N_1(\omega)$, such that $\gamma_T(u) \in \bar{B}_{\frac{1}{2}r_{\gamma(S)}}(\gamma(S))$ for all $T > N_1$ and all $u \in (S - \delta, S)$. Because the γ_T's are uniformly continuous on $(S - \delta, S)$, and $\bar{B}_{r_{\gamma(S)}}(\gamma(S))$ is compact, the γ_T's are also defined in S for all $T > N_1$. By the equiLipschitz property of the γ_T, we also have that $\lim_{T\uparrow\infty} \gamma_T(S) = \gamma(S)$. We are in the same situation as in Step 1 now. From this, it is clear that we can extend the interval on which statement (4.58) is valid, which contradicts the definition of S. Hence, $S(\omega) = \nu(\omega)$, which proves Step 2.

Step 3. The rest of the proof of the proposition.

1. For all $s \in [0, \nu(\omega))$, statement (4.58) is valid on the interval $[0, s]$. Consequently, $\liminf_{T\uparrow\infty} \frac{\nu_T}{T} \geq s$, thus also $\liminf_{T\uparrow\infty} \frac{\nu_T}{T} \geq \nu$. This proves part 1. of the proposition.

2. *We are going to show that ν is a stopping time.*
Before proving this, we show that the radius r_{q_0}, which we used in the proof of Step 1, is a measurable function w.r.t. $\mathcal{B}_0$. We do this by showing that we can choose r_p, such that it is a continuous function of p. Define the function $\rho : M \to \mathbb{R} \cup \{\infty\}$ by:

$$\rho(p) = \sup\{r \in \mathbb{R}_+ |\ \bar{B}_r(p) \text{ is compact, and contained in an open set, which is a normal neigbourhood of each of its points}\} \quad (4.67)$$

From Theorem I.9.9., Helgason, we known that $\rho(p) > 0$ for all $p \in M$. If $d(p,q) > \varepsilon$, then $\rho(q) > \rho(p) - \varepsilon$. Let $Y \in \mathbb{R}^+$. Then define:

$$r_p = \frac{1}{2}\min(\rho(p), Y).$$

This function is obviously continuous on M. Furthermore, $\bar{B}_{r_p}(p)$ is compact and contained in an open set which is a normal neighbourhood of its points. Consequently, $r_{q_0(\omega)} : \Omega \to \mathbb{R}$ is measurable w.r.t. $\mathcal{B}_0$., proving our claim.

In the first part of Step 1, (see equation (4.61)), we constructed a number $s(\omega)$, such that

$$\int_0^u \sum_{i=1}^m |\theta_i(\tau)| d\tau < \frac{r_{q_0}}{2K}, \text{ if } 0 \le u \le s(\omega). \quad (4.68)$$

(Here, K is the number, defined in equation (4.60)). This was sufficient to prove that on the interval $[0, s(\omega)]$ statement (4.58) was valid. We'll make a stopping time out of this condition by defining:

$$s_0(\omega) = \inf\{u \in \mathbb{R}_+ | \int_0^u \sum_{i=1}^m |\theta_i(\tau)| d\tau - \frac{r_{q_0}}{2K} \ge 0\}. \quad (4.69)$$

Indeed, s_0 is a stopping time, because

$$\{s_0 > w\} = \{\int_0^w \sum_{i=1}^m |\theta_i(\tau)| d\tau - \frac{r_{q_0}}{2K} < 0\} \in \mathcal{B}_w.$$

(Note that the stochastic variables $\int_v^u \sum_{i=1}^m |\theta_i(\tau)| d\tau \quad (0 \le v < u)$ are measurable w.r.t. $\mathcal{B}_u$ for all $0 \le v < u$.)
Statement (4.58) is valid on the interval $[0, s_0(\omega)]$, hence we have:

$$\lim_{T \uparrow \infty} \gamma_T(u \wedge s_0 \wedge \frac{\nu_T}{T}) = \gamma(u \wedge s_0).$$

Because $\gamma_T(u \wedge s_0 \wedge \frac{\nu_T}{T})$ is measurable w.r.t. $\mathcal{B}_u$, also $\gamma(u \wedge s_0)$ is measurable w.r.t. $\mathcal{B}_u$. In order to prove that ν is a stopping time, we have to show that the set $\{\nu > u_0\}$ is an element of $\mathcal{B}_{u_0}$ for all $u_0 > 0$. We already know that $s_0 > u_0$ implies that $\nu > u_0$. If also $\nu > u_0$ would imply $s_0 > u_0$, then the proof would have been complete, because s_0 is a stopping time. But the last implication need not to be

true: condition (4.68) was only sufficient, not necessary for statement (4.58) to be valid on interval $[0, s]$. We overcome this difficulty by defining a double sequence of stopping times.
Fix $u_0 > 0$. For any $n \in \mathbb{N}$, we recursively define the stochastic variables:

$$s_{1,n} = s_0 \tag{4.70}$$

$$s_{k+1,n} = s_{k,n} \text{ if } s_{k,n} \leq \frac{ku_0}{n}; \tag{4.71}$$

$$\text{if } s_{k,n} > \frac{ku_0}{n} \text{ then}$$

$$s_{k+1,n} = \inf\{u \geq \frac{ku_0}{n} | \int_{\frac{ku_0}{n}}^{u} \sum_{i=1}^{m} |\theta_i(\tau)| d\tau - \frac{r_{\gamma(\frac{ku_0}{n} \wedge s_{k,n})}}{2K} \geq 0\}. \tag{4.72}$$

Note that $s_{k,n} > \frac{ku_0}{n}$ implies that $s_{k+1,n} > \frac{ku_0}{n}$. By induction, we show that $s_{k,n}$ is a stopping time for all $k \in \mathbb{N}$, and that $\gamma(u \wedge s_{k,n})$ is measurable w.r.t. $\mathcal{B}_u$ for all $u \in \mathbb{R}_+$. Only the induction step has to be proved:

$$\text{if } w \leq \frac{ku_0}{n}, \text{ then } \{s_{k+1,n} \leq w\} = \{s_{k,n} \leq w\} \in \mathcal{B}_w \text{ and if } w > \frac{ku_0}{n} \text{ then}$$

$$\{s_{k+1,n} > w\} = \{s_{k,n} > \frac{ku_0}{n}\} \cap \{\int_{\frac{ku_0}{n}}^{w} \sum_{i=1}^{m} |\theta_i(\tau)| d\tau - \frac{r_{\gamma(\frac{ku_0}{n} \wedge s_{k,n})}}{2K} < 0\} \in \mathcal{B}_w.$$

Just as in the case of s_0 one can prove now that $\gamma_T(u \wedge s_{k+1,n} \wedge \frac{\nu_T}{T})$ is measurable w.r.t. $\mathcal{B}_u$, and converges to $\gamma(u \wedge s_{k+1,n})$, which consequently must then also be measurable w.r.t. $\mathcal{B}_u$.

Now, we are able to prove that ν is a stopping time.
If $\nu > u_0$, then for some $\varepsilon > 0$, the curve $\{\gamma(v) | \, v \in [0, u_0 + \varepsilon]\}$ forms a compact part of M. The function $p \mapsto r_p$ has a positive minimum on this curve. The function $v \mapsto \int_0^v |\theta_i(\tau)| d\tau$ is uniformly continuous on $[0, u_0 + \varepsilon]$. Hence, there is an $n \in \mathbb{N}$, such that $s_{k,n} > \frac{ku_0}{n}$ for all k, $1 \leq k \leq n$. Consequently, there is an $n \in \mathbb{N}$, such that $s_{n,n} > u_0$. On the other hand, if there is an $n \in \mathbb{N}$, such that $s_{n,n} > u_0$, then automatically $\nu > u_0$. Hence,

$$\{\nu > u_0\} = \bigcup_{n \in \mathbb{N}} \{s_{n,n} > u_0\} \in \mathcal{B}_{u_0}.$$

This proves the first statement of this part 2. of proposition 4.10.
Next, the second statement. If M is complete, all geodesics are defined on the whole $\mathbb{R}$. Likewise, also the polygons are defined on $\mathbb{R}_+$, and, consequently, $\nu_T = \infty$. Suppose that $\nu < \infty$. As the $(\gamma_T)_{T \in \mathbb{N}}$'s uniformly converge to γ on $[0, \nu)$, the sequence $(\gamma_T(\nu))_{T \in \mathbb{N}}$ is Cauchy. Because a geodesically complete manifold is also metrically complete (Helgason, Theorem 10.3), this sequence converges to a point $p \in M$. The same reasoning as in (a) shows that the differential equation (4.50) has also

a solution on an interval $[0, \nu + \varepsilon)$, which contradicts the maximality of γ, i.e. the definition of ν. So $\nu = \infty$ (with probability 1).
In case $\nu < \infty$, we now know that M can't be complete. We can form the completion $\bar{M}$ of M as a metric space. In $\bar{M}$, we can extend γ in a unique way to a continuous curve, defined on the segment $[0, \nu]$, because equation (4.63) shows that γ is uniformly continuous on $[0, \nu)$. In this case, $\gamma(\nu)$ is an element of ∂M, the boundary of M in $\bar{M}$.

3. The minimum of a finite sequence of stopping times is again a stopping time, hence $\mu_T = R \wedge \nu \wedge \frac{\nu_T}{T}$ and $\mu = R \wedge \nu$ are stopping times, and $\gamma_T(u \wedge \mu_T)$ is $\mathcal{B}_u$-measurable for all u.
We prove uniform convergence of $\gamma_T(u \wedge \mu_T)$ to $\gamma(u \wedge \nu)$ with probability 1.
Consider the functions:

$$f_T : [0, R] \cap [0, \nu) \to \mathbb{R} \; ; \; f_T(u) = d(\gamma_T(u \wedge \mu_T), \gamma(u \wedge \nu)) \tag{4.73}$$

For all $u, v \in [0, R] \cap [0, \nu)$, we have that:

$$\begin{aligned} d(\gamma_T(u), \gamma(u)) &\leq d(\gamma_T(u), q_0) + d(\gamma(u), q_0) \leq & (4.74) \\ &\leq d(\gamma_T(0), q_0) + \frac{1}{T} \sum_{t=0}^{[RT]+1} || \sum_{i=1}^{m} \theta_i(\frac{t}{T}) X_i || + \\ &+ \int_0^R || \sum_{i=1}^{m} \theta_i(v) X_i || dv., & (4.75) \end{aligned}$$

and also

$$|d(\gamma_T(u), \gamma(u)) - d(\gamma_T(v), \gamma(v))| \leq d(\gamma_T(u), \gamma_T(v)) + d(\gamma(u), \gamma(v)). \tag{4.76}$$

These inequalities, together with inequality (4.63), show that the functions f_T are equiLipschitz, and uniformly bounded. Because of the Lipschitz property, they can uniquely be extended to a set of uniformly bounded functions on $[0, R]$, which are also equiLipschitz. We already know that the $(f_T)_{T \in \mathbb{N}}$'s converge to 0 on $[0, R] \cap [0, \nu)$ pointwise, but by Arzela-Ascoli, we now also know that the $(f_T)_{T \in \mathbb{N}}$'s have to converge to 0 on $[0, R]$ uniformly with probability 1.
Next, we prove that $\gamma_T(u \wedge \mu_T)$ converges to $\gamma(u \wedge \nu)$, uniformly in $u \in [0, R]$, also in mean square sense.
First, we show that the $(f_T)_{T \in \mathbb{N}}$'s are uniformly bounded and continuous in mean square sense, i.e.

$$\exists P \forall T \in \mathbb{N} \quad \boldsymbol{E} f_T^2 < P, \text{ and } \forall T \forall u \in [0, R] \quad \lim_{v \to u} \boldsymbol{E} |f_T(u) - f_T(v)|^2 = 0.$$

Then, we prove convergence in mean square by the majorized convergence theorem of Lebesgue. We have to restate a proof of that theorem, because we want to prove *uniform* convergence in mean square.
Because $\boldsymbol{E}|\theta_i(u+h) - \theta_i(u)|^2 = 2\tau^2(1 - e^{-\beta_i |h|})$, all θ_i are mean- square- continuous

functions on $[0,R]$, and $u \mapsto ||\sum_{i=1}^m \theta_i(u,\omega)X_i||$ is integrable on $[0,R]$ in the mean square sense (see e.g. Yazwinski, Theorem 3.7, Corollary 2). It is also easily verified that $\frac{1}{T}\sum_{t=1}^{[RT]} ||\sum_{i=1}^m \theta_i(\frac{t}{T})X_i||$ converges to $\int_0^R ||\sum_{i=1}^m \theta_i(u)X_i||$ in mean square sense. So, for all $\varepsilon > 0$ there is an N, such that for all $T > N$ we have:

$$\boldsymbol{E}(\int_0^R ||\sum_{i=1}^m \theta_i(u)X_i||du)^2 + \varepsilon \geq \boldsymbol{E}(\frac{1}{T}\sum_{t=1}^{[RT]} ||\sum_{i=1}^m \theta_i(\frac{t}{T})X_i||)^2. \qquad (4.77)$$

From this and the inequalities (4.75, 4.76), it follows that the $(f_T)_{T\in\mathbb{N}}$'s are uniformly bounded by a certain number $P > 0$ and are, moreover, uniformly continuous in mean square sense.
Next, we define:

$$h_T = (\frac{1}{T}\sum_{t=1}^{[RT]} ||\sum_{i=1}^m \theta_i(\frac{t}{T})X_i|| + \int_0^R ||\sum_{i=1}^m \theta_i(u)X_i||)^2 - f_T(u)^2$$

and

$$h = 4(\int_0^R ||\sum_{i=1}^m \theta_i(u)X_i||du)^2.$$

Then, by (4.75),

$$h_T \geq 0 \ \forall T \in \mathbb{N} \text{ and } \lim_{T\uparrow\infty} h_T(u,\omega) = h(u,\omega), \text{ a.s. uniform in } u \in [0,R].$$

Also $u \mapsto \boldsymbol{E}h_T(u)$ is continuous by mean square continuity of f_T.
The increasing sequence $(g_T)_{T\in\mathbb{N}}$, $g_T := inf_{k\geq T} h_T$ is also well defined, measurable and converges to h a.s. and uniform in u. Let $g(\omega) = \sum_{j=1}^p a_i 1_{A_i}(\omega)$ be a measurable function with a finite number of values, such that $0 \leq g \leq h$. Let $\varepsilon > 0$. We define the measurable set $E_T = \{\omega \in \Omega | g_T(u) \geq (1-\varepsilon)g \ \forall u \in [0,R] \cap \mathbb{Q}\}$. By uniform convergence a.e., the sequence of these sets E_T is increasing, and its union is Ω. Consequently, for all measurable sets A, $\lim_{T\uparrow\infty} P(A \cap E_T) = P(A)$. Now, it follows that

$$h_T(u) \geq g_T(u) \geq (1-\varepsilon)g1_{E_T} = (1-\varepsilon)\sum_{j=1}^p a_i 1_{A_i\cap E_T}, \ \forall u \in [0,R] \cap \mathbb{Q}.$$

Hence, using (4.77), there is an N, such that for all $T > N$

$$\boldsymbol{E}h + 4\varepsilon \geq \boldsymbol{E}h_T(u) \geq (1-\varepsilon)\sum_{j=1}^p a_i P(A_i \cap E_T) \ \forall u \in [0,R] \cap \mathbb{Q}.$$

The last line holds for all $u \in [0,R]$ since $u \mapsto \boldsymbol{E}h_T(u)$ is continuous. As $\boldsymbol{E}h$ is the supremum of the expectations of all measurable functions smaller than h with a finite number of values, the proof of part 3 of the proposition is now easily concluded. □

If the coefficient processes in the sequence are stopped at their running times, or earlier, then the sequence, after re-indexation, satisfies an equislowness property. We write this formally down below.

Let $(q_{t,T}, d_{t,T})_{t\in\mathbf{Z}_+, T\in\mathbf{N}}$ be a sequence of coefficient processes, modelled according to the Main Model Sequence in equations (4.44-4.46). Let $R \in \mathbf{R}^+$ be the relative consideration time of our study of this sequence of processes, i.e. we shall only consider the first $[RT]$ elements of process $(q_{t,T}, d_{t,T})_{t\in\mathbf{Z}_+}$. Assume that these processes are of independent directional variation until their running times ν_T. Then $[RT]\wedge\nu_T$ is a stopping time of the process $(q_{t,T}, d_{t,T})_{t\in\mathbf{Z}_+}$ smaller than ν_T. Accordingly, this process is also of independent directional variation until $[RT] \wedge \nu_T$.

Stopping and re-indexation

We define the sequence of stochastic processes $(\tilde{q}_{t,RT}, \tilde{d}_{t,RT})_{0\le t\le RT}$ (with sequence of σ-algebras) by:

$$\begin{aligned} \tilde{q}_{t,RT} &= q_{t\wedge[RT]\wedge\nu_T,T} \\ \tilde{d}_{t,RT} &= d_{t\wedge[RT]\wedge\nu_T,T} \qquad (0 \le t \le RT) \qquad (4.78) \\ \tilde{\mathcal{B}}_{t,RT} &= \mathcal{B}_{t,T} \qquad (0 \le t \le RT) \qquad (4.79) \end{aligned}$$

These processes are just the *stopped* processes $(q_{t,T}, d_{t,T})$, stopped at time $[RT] \wedge \nu_T$, and re-indexed in order to let the second index indicate the number of path points, taken into consideration. Also these stopped processes are of independent directional variation until $[RT]\wedge\nu_T$. As an immediate consequence of the previous proposition (4.10), we have:

Corollary 4.12 1. The $(\tilde{q}_{t,S})_{0\le t\le S; S\in RN}$'s vary equislowly with the increasing number of path points, i.e. there is a number $C(\omega)$, such that

$$d(\tilde{q}_{t+1,S}, \tilde{q}_{t,S}) \le \frac{C(\omega)}{S};$$

2. If $\nu > R$, then for all $u \in [0,1]$ we have w.p. 1:

$$\lim_{S\uparrow\infty} \tilde{q}_{[uS],S} = \gamma(uR) \in M;$$

3. If $\nu \le R$, then there is a point $p = \frac{\nu}{R}$ in $[0,1]$ such that w.p.1:

$$\lim_{S\uparrow\infty} \tilde{q}_{[pS],S} = \gamma(\nu) \in \partial M.$$

Proof: To see part 1. of corollary 4.12, just use equation (4.63), derived in Step 1. of the proof of proposition 4.10. To prove part 2. and part 3. of the corollary, just use part 3. of the proposition. □

4.5 A conditional spectral density

We consider a sequence of univariate time-varying AR(n) processes with stochastic coefficients:

$$y_{t,T} = a_{1,t,T}y_{t-1,T} + \cdots + a_{n,t,T}y_{t-n,T} + \sigma_{t,T}\epsilon_{t,T} \quad (t \in \mathbf{Z}_+,\ T \in \mathbb{N}) \tag{4.80}$$

where $\sigma_{t,T} > 0$ for all $t \in \mathbf{Z}_+, T \in \mathbb{N}$, and the $(\epsilon_{t,T})_{t\in\mathbf{Z}_+}, T \in \mathbb{N}$ are Gaussian white noise sequences with variances equal to one, independent of $y_{-1,T}, \cdots, y_{-n,T}$ and also independent of the stochastic process, generating the coefficients.

About the coefficients (or *instantaneous associated polynomials*, as we called these earlier) and pre-initial observations $y_{-n,T}, \cdots, y_{-1,T}$, we make the following **Assumptions:**

1. The instantaneous associated polynomials

$$q_{t,T} = L_{n,t,T} = \frac{1}{\sigma_{t,T}}(x^n - a_{1,t,T}x^{n-1} \ldots - a_{n,t,T})$$

are walking stochastically in U_n, the set of stable polynomials of degree n with positive highest coefficient, which is embedded in $\mathbb{R}^{n+1}$ in the "natural" way :

$$U_n \ni e^{-r}(x^n + a_1x^{n-1} + \cdots a_{n-1}x + a_n) \mapsto \begin{pmatrix} a_1 \\ \vdots \\ a_n \\ r \end{pmatrix} \in \mathbb{R}^{n+1}.$$

2. These instantaneous associated polynomials are walking in a Riemannian manifold M that is a submanifold of U_n. Note that we now have two distance measures on M: its own Riemannian distance measure $\boldsymbol{d}$ *and* the Euclidean distance $||\ ||$, induced by the "natural" embedding in $\mathbb{R}^{n+1}$. The first was used in proposition 4.10 and corollary 4.12, the latter in proposition 4.4 and proposition 4.6. On compact subsets of M the two distances are equivalent, as shown in lemma 2.2 in chapter two.

3. The instantaneous associated polynomials are generated, according to the Main Model (4.44-4.46); the polygonal processes are of independent directional variation until their running times.

4. The initial instantaneous associated polynomials $q_{0,T}$ are all equal to a stochastic polynomial $q_0 : \Omega \to M$, independent of the OUV(β_i, τ_i) processes, used in the Main Model.

5. The pre-initial observations $y_{-n,T}, \cdots, y_{-1,T}$ are independent of these OUV(β_i, τ_i) processes. Conditional on q_0, these observations have a zero mean Gaussian distribution with variance as if they belong to a time-invariant AR(n) process with associated polynomial q_0. Hence, conditionally on q_0, but independently from the OUV(β_i, τ_i) processes, the processes $(y_{\cdot,T})$, $T \in \mathbb{N}$ can and will be extended to the infinite past. (This is just a technicallity of the same kind as used in section 4.3).

6. In order to derive also properties on the process after the asymptotic relative running time, we make an assumption on the definition of the $q_{t+1,T}$ if $\exp_{q_{t,T}}(d_{t,T})$ is not well-defined. Let the $\mathcal{B}_{t,T}$-measurable variables $h_{t+1,T}$ of equation (4.46) satisfy:

$$d(q_{t,T}, h_{t+1,T}) \leq ||d_{t,T}||_{q_{t,T}}, \text{ a.s., if } \nu_T \leq t.$$

In combination with equation (4.46), this assumption implies that

$$d(q_{t,T}, q_{t+1,T}) \leq ||d_{t,T}||_{q_{t,T}} \text{ a.s..}$$

Let $R \in \mathbb{R}^+$ be the relative consideration time for which we study the sequence of processes $(y_{t,T})_{0 \leq t \leq RT,\ T \in \mathbb{N}}$. In order to apply concepts as BIBO-stability and Dahlhaus spectral density, it would have been necessary to re-index this sequence. We did so in Corollary (4.12). We shall omit that now.

Remind that $\mathcal{B}_R$ is the smallest σ-algebra in $\mathcal{A}$ with respect to which q_0, and $\theta_i(u)$ $(1 \leq i \leq m,\ 0 \leq s \leq R)$ are measurable. The σ-algebras $\mathcal{B}_{t,T}$ were defined in (4.44-4.46).

Almost trivial is the following **Statement.** Let $T, t \in \mathbb{N}$ with $t \leq RT$.
For all polynomial functions $h : \mathbb{R}^{t+1+n} \to \mathbb{R}$, we have:

$$E^{\mathcal{B}_R} h(y_{-n,T}, \cdots, y_{t,T}) = E^{\mathcal{B}_{t,T}} h(y_{-n,T}, \cdots, y_{t,T}).$$

The conditional distribution of $y_{-n,T}, \cdots, y_{t,T}$, given $\mathcal{B}_R$, is the same as the conditional (Gaussian) distribution of $y_{-n,T}, \cdots, y_{t,T}$, given $\mathcal{B}_{t,T}$.
Proof: We can rewrite $h(y_{-n,T}, \cdots, y_{t,T})$ as
$g(y_{-n,T}, \cdots, y_{-1,T}, q_0, \lambda_{0,T}, \cdots, \lambda_{t-1,T}, \epsilon_{0,T}, \cdots, \epsilon_{t,T})$ for some function g that is polynomial in $\epsilon_{t,T}$ and $y_{\cdot,T}$, and where $\lambda_{s,T} \in \mathcal{B}_{s,T}$ was defined in equation (4.48). The terms $\epsilon_{s,T}$ and the pre-initial observations $y_{-n,T}, \cdots, y_{-1,T}$ are supposed to be independent not only from $\lambda_{1,T}, \cdots, \lambda_{t-1,T}$, but also from the OUV$(\beta_i, \tau_i)$ processes θ_i $(1 \leq i \leq m)$, even conditional on q_0. The first statement is now immediate. The second statement follows from the first. □

Conditional on $\mathcal{B}_R$, we associate with the process $(y_{t,T})_{0 \leq t \leq RT}$, a system $\Sigma_T^{\mathcal{B}_R}$ with time-varying system matrices $(C_{t,T}, A_{t,T}, B_{t,T})$, where

$$C_{t,T} = \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix}$$

$$A_{t,T} = \begin{pmatrix} a_{1,t+1,T} & a_{2,t+1,T} & \cdots & a_{n-1,t+1,T} & a_{n,t+1,T} \\ 1 & 0 \cdots & 0 & 0 & \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}$$

$$B_{t,T} = \begin{pmatrix} \sigma_{t+1,T} \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \qquad (0 \leq t \leq RT-1). \tag{4.81}$$

If one defines $X_{t,T} = \begin{pmatrix} y_t & y_{t-1} & \cdots & y_{t-n} \end{pmatrix}$, one can write, conditional on $\mathcal{B}_R$,

$$\Sigma_T^{\mathcal{B}_R} : \left\{ \begin{array}{lcl} X_{t+1,T} & = & A_{t,T}X_{t,T} + B_{t,T}\epsilon_{t,T} \\ y_{t,T} & = & C_{t,T}X_{t,T} \end{array} \right\} \quad (0 \leq t \leq RT). \tag{4.82}$$

Let $A \in \mathcal{B}_R$ be an event with positive probability. We shall call the sequence of processes $(y_{t,T})_{0 \leq t \leq RT}, T \in \mathbb{N}$ conditionally BIBO stable on the event A if on this event the corresponding sequence of systems $(\Sigma_T^{\mathcal{B}_R})_{T \in \mathbb{N}}$ is BIBO stable.

Let $A \in \mathcal{B}_R$ be an event with positive probability. We shall say that the sequence of processes $(y_{t,T})_{0 \leq t \leq RT}$, $T \in \mathbb{N}$ admits the conditional time-varying spectral density $f(u, w|\theta)$ on the event A if on this event the following limits exist:

$$f(u,w|\theta) = \lim_{T \to \infty} \sum_{s=-\infty}^{\infty} w^s \boldsymbol{Cov}^{\mathcal{B}_r}(y_{[uT+\frac{r}{2}+\frac{s}{2}],T}, y_{[uT+\frac{r}{2}-\frac{s}{2}],T}), \tag{4.83}$$

uniformly in $u \in (-\infty, R]$, and $w \in \mathbf{T}$ (i.e. $w = e^{i\psi}$, $\psi \in [-\pi, \pi]$) for all $r \in \mathbf{Z}$. (the summation is taken over all s, such that $\boldsymbol{Cov}^{\mathcal{B}_R}(y_{[uT+\frac{r}{2}+\frac{s}{2}],T}, y_{[uT+\frac{r}{2}-\frac{s}{2}],T})$ is defined).

Theorem 4.13 Under the Assumptions 1,2,3,4,5, we have:

1. The sequence of time-varying AR processes $(y_{t,T})_{0 \leq t \leq RT}$ is conditionally BIBO stable, and admits the conditional Dahlhaus spectral density on the event $\{\nu > R\}$, i.e. on the event that the asymptotic relative running time is larger than the relative consideration time.

2. If the manifold M is complete, then the sequence of processes $(y_{t,T})_{0 \leq t \leq RT}$ is conditionally BIBO stable, and admits the conditional Dahlhaus spectral density with probability 1.

3. Suppose that assumption 6 is valid. If the closure of M, as part of $\mathbb{R}^{n+1}$, is compact and lies in U_n, then the sequence of processes $(y_{t,T})_{0 \leq t \leq RT}$ is conditionally BIBO stable with probability 1, and it admits the conditional Dahlhaus spectral density on the event $\{\nu \geq R\}$.

4. Suppose that assumption 6 is valid. If M coincides with U_n, and the Euclidean distance is smaller than a multiple of the Riemannian distance everywhere, then the sequence of processes $(y_{t,T})_{0\leq t\leq RT}$ is conditionally BIBO-stable, and admits the conditional Dahlhaus spectral density on the event $\{\nu > R\}$ and on this event only.

Proof:

1. On the event that $\nu > R$, we have that $\nu_T > RT$, if $T > N$ for some $N(\omega) \in \mathbb{N}$, hence $\tilde{q}_{t,RT} = q_{t,T}$ for all $0 \leq t \leq RT$ and $T > N$. Here, $\tilde{q}_{t,RT}$ was defined in (4.78). By corollary 4.12, we know that the coefficients $(\tilde{q}_{t,RT})_{0\leq t\leq RT,\ RT\in\mathbb{N}}$ are equislowly varying in the sense of the metric. If $\nu > R$, then there is only one limit curve completely lying in U_n; this limit curve on $[0, R]$ is compact, hence the coefficients lie in a compact subset of U_n. By lemma 2.2 in chapter two, the coefficients are also bounded, equislowly varying, and have one limit curve, in Euclidean sense. Proposition 4.4 and proposition 4.6 now show that the sequence of processes $(y_{t,T})_{0\leq t\leq RT;\ T\in\mathbb{N}}$ is conditionally BIBO stable, and admits the Dahlhaus time-varying spectral density.

2. If M is complete, the asymptotic relative running time is infinite; $\nu = \infty$ (see proposition (4.10). The second statement of the theorem now follows from the first.

3. If the closure of M in $\mathbb{R}^{n+1}$ is compact in U_n, then the coefficients lie in a compact subset of U_n, whatever the value of ν. The definition of the $d_{t,T}$ in equation (4.45) and assumption 6, on the choice of the $q_{t+1,T}$ after the running time ν_T, make clear that the coefficients $(q_{t,T})_{0\leq t\leq \nu T}$, $T \in \mathbb{N}$ are equislowly varying, and are bounded in metrical sense. Again, according to lemma 2.2 of chapter two, they are equislowly varying, and bounded in Euclidean sense. As all the limit curves of the coefficients lie in U_n (in metrical, but also in Euclidean sense), the sequence of systems is conditionally BIBO stable, according to proposition 4.4. If $\nu \geq R$, there is only one limit curve, hence the sequence of systems admits Dahlhaus' spectral density, according to 4.6.

4. If, for some number C, we have that $||p - q|| \leq C\boldsymbol{d}(p, q)$ for all $p, q \in M$, then equislowness, boundedness and convergence in metrical sense imply equislowness, boundedness and convergence in Euclidean sense. Even convergence to a point in the completion of M implies convergence to a point in the boundary of U_n in $\mathbb{R}^{n+1}$. We show that the sequence of processes $(y_{t,T})_{0\leq t\leq \nu T}$, $T \in \mathbb{N}$, defined conditional on the event $\{\nu \leq R\}$, can't be BIBO stable. On this event, then also the sequence of processes $(y_{t,T})_{0\leq t\leq RT}, T \in \mathbb{N}$ can't be conditionally BIBO stable, according to the definition of BIBO stability. By the definition of the $d_{t,T}$ in equation (4.45) and assumption 6, the coefficients $(q_{t,T})_{0\leq t\leq \nu T}$, $T \in \mathbb{N}$ vary equislowly, and are bounded in metrical sense. The limit curve is $\gamma : [0, \nu] \to \tilde{M}$, where $\tilde{M}$ is the

completion of M and $\gamma(\nu) \notin M$. Then, also in Euclidean sense the coefficients have this limit curve, and $\gamma(\nu) \notin U_n$. According to proposition 4.4, the sequence $(y_{t,T})_{0 \leq t \leq \nu T}$, $T \in \mathbb{N}$ is not conditionally BIBO stable, and according to proposition 4.6, it does not admit the Dahlhaus spectral density. Then, also the sequence $(y_{t,T})_{0 \leq t \leq RT}, T \in \mathbb{N}$ can't admit this density. □.

Corollary 4.14 Assumptions 1,2,3,4,5 as in the Theorem. On the event that $\nu > R$ and conditional on $\mathcal{B}_R$, the autocovariances of the processes $(y_{t,T})_{0 \leq t \leq RT}$ are asymptotically:

$$\lim_{T \uparrow \infty} \boldsymbol{Cov}^{\mathcal{B}_R}(y_{[uT-j],T}, y_{[uT-l,T]}) = \int_{\mathbf{I}} \bar{w}^{l-j} f(u,w|\theta) dm(w)$$
$$\text{for all } l,j \in \mathbf{Z};\ \text{for all } u \in [0,R]. \quad (4.84)$$

Proof: Just as the proof of corollary 4.7. □

4.6 The use of the spectral density, an heuristic discussion

In this section, we shall explain what our aim was in constructing this concept of a conditional time-varying spectral density, and how far we have got in achieving our goals. We give some indications how one could use this density.
First of all, by using corollary 4.14, we have been able to prove asymptotic properties of zero-curvature models, which makes the role of the metric clear. These results will be presented in chapters five and six.
Secondly. The usual stochastic coefficient models have in common with the models proposed in chapter three that it will be necessary for a useful application to estimate the hyperparameters β_i, τ_i of the (smoothly integrated) random walk, or the OUV-process. This is already not an easy question. But on top of this we added a lot more structure to be specified: the manifold with metric and the direction space with eigenspaces of the symmetric tensor field. With the extra structure, the number of hyperparameters involved has increased, and becomes comparable to the number of parameters in parametric deterministic models. Dahlhaus has constructed and analyzed a method to fit these deterministic models in practical situations. A method that is closely related to maximum likelihood estimation. Our idea is that it must be possible to use his methods and analyses for the estimation of the *extra* structure we added to the models in chapter three. The idea is to estimate the extra structure after the choice of β_i, τ_i has been made. In other words, the estimation will be conditional on our a priori choice of OUV-processes β_i, τ_i. Let us explain the method we propose. It uses ideas and results from Dahlhaus. The elaboration of the method remains to be done; this section has an heuristic character.
Let $y_0, \cdots, y_T$ be a series of (univariate) observations, supposed to be the realization of the Tth element of a sequence of finite Gaussian equislowly time-varying AR(l) processes $(y_{t,T})_{0 \leq t \leq T}$, $T \in \mathbb{N}$, for some order l, with probability densities $p^*_T(z)$. Assume that the sequence admits a time-varying spectral density $f^*(u,w)$. We are going to fit a model of a sequence of finite Gaussian equislowly time-varying AR(n) processes that admit a time-varying spectral density. First, we discuss the case that the model belongs to a class of parametric deterministic models. The probability densities of a model in the class with parameter ζ will be denoted as $p_{\zeta,T}(z)$ and its spectral density as $f_\zeta(u,w)$. We have to choose a criterion for optimal fitting of a model in the class to the given sequence of processes. The sequences of processes, on which proposition 4.6 applies, can be regarded as *doubly indexed locally stationary processes*, as defined by Dahlhaus. In order to fit a model of a class of locally stationary doubly indexed processes to a given locally stationary process, Dahlhaus proposes to minimize the asymptotic Kullback-Leibler divergence between the probability distributions of a model in the class and the probability distributions of the given sequence of processes. Hence, he proposes to minimize

$$D(p_\zeta, p^*) = \lim_{T \uparrow \infty} \frac{-1}{T+1} \boldsymbol{E}_{p^*_T} \log\left(\frac{p_{\zeta,T}}{p^*_T}\right). \tag{4.85}$$

He succeeds in writing this divergence by means of the spectral densities as

$$D(p_\zeta, p^*) = \frac{1}{2}\int_0^1 \int_{\mathbf{I}} \{\log(\frac{f_\zeta(u,w)}{f^*(u,w)}) + \frac{f^*(u,w)}{f_\zeta(u,w)} - 1\} dm(w) du. \tag{4.86}$$

Minimization of this criterion w.r.t. ζ is the same as minimizing the criterion

$$\mathcal{J}(\zeta) := \frac{1}{2}\int_0^1 \int_{\mathbf{I}} \{\log(f_\zeta(u,w)) + \frac{f^*(u,w)}{f_\zeta(u,w)}\} dm(w) du. \tag{4.87}$$

It is not possible to minimize $\mathcal{J}$ immediately, because the spectral density $f^*(u,w)$ of the observed sequence of processes is unknown. Dahlhaus proposes to estimate this by a "local version of the periodogram",

$$I_N(u,w) = \frac{|\sum_{i=0}^{N-1} h(\frac{s}{N}) \bar{w}^s y_{[uT]-\frac{N}{2}+s+1,T}|^2}{\sum_{s=0}^{N-1} h(\frac{s}{N})^2} \tag{4.88}$$

where N is an even number and $h : \mathbf{R} \to \mathbf{R}$ is a continuous function which is zero outside the interval $[0,1)$. Then his new criterion becomes:

$$\mathcal{J}_T(\zeta) = \frac{1}{2}\frac{1}{S}\sum_{j=1}^{S} \int_{\mathbf{I}} \{\log(f_\zeta(u_j,w) + \frac{I_N(u_j,w)}{f_\zeta(u_j,w)}\} dm(w). \quad (u_j = \frac{j}{S}) \tag{4.89}$$

He succeeds in proving that, under certain conditions, a minimizing $\hat{\zeta}_T$ of $\mathcal{J}_T(\zeta)$ converges to a minimizing ζ_0 of $\mathcal{J}(\zeta)$, and even succeeds in determining the asymptotic distribution of $\sqrt{T}(\hat{\zeta}_T - \zeta_0)$.

Now, we want to fit a class of Main Model Sequences, as defined earlier. The Model Sequences in the class satisfy the Assumptions 1,2,3,4,5 of theorem 4.13. As we said above, we have chosen the hyperparameters β_i, τ_i of the OUV-processes θ a priori. Every realization $(\theta(u))_{u\in[0,1]}$ (or $(\theta(0), \theta(\frac{1}{T}), \cdots, \theta(1))$) of such an OUV-process can be seen as a *parameter* of the model sequence (or model number T, respectively) belonging to the parameter space Par, on which we have laid our prior OUV distribution, determined by β_i, τ_i. A Model Sequence in the class depends furthermore on the extra structure, we referred to above, and which we summarize in the hyperparameter ζ. The corresponding asymptotic relative running time will be indicated by ν_ζ; the relative consideration time will be one, as in the beginning of chapter four. On the event $\{\nu_\zeta > 1\}$, the model sequence admits the spectral density $f_\zeta(u,w|\theta)$. The joint probability density on the product of sample space and parameter space for process number T of the model sequence will be denoted by $p_{\zeta,T}(z,\theta)$. According to Bayes'rule, we have:

$$p_{\zeta,T}(z,\theta) = p_{\zeta,T}(z|\theta) p_T(\theta) \tag{4.90}$$

where $p_T(\theta) = p(\theta(0), \theta(\frac{1}{T}), \cdots, \theta(1))$ is a marginal probability density of the OUV process θ and $p_{\zeta,T}(z|\theta)$ the density of the conditional Gaussian distribution, given $\mathcal{B}_{T,T}$. We have seen before that this distribution is equal to the conditional distribution, given the whole σ-algebra $\mathcal{B}_1$. We want to compare the given sequence of processes with the model sequences of the class on these products of sample spaces and parameter spaces. As the

given processes do not depend on our a priori chosen OUV process, the corresponding density of process number T on sample space times parameter space will be

$$p_T^*(z,\theta) = p_T^*(z)p_T(\theta). \tag{4.91}$$

A logical measure of fit will be the Kullback-Leibler divergence over both variables z and θ:

$$\begin{aligned} D_T(p_\zeta, p^*, Sam \times Par) &= \frac{-1}{T+1}\int\int p_T^*(z,\theta)\log(\frac{p_{\zeta,T}(z,\theta)}{p_T^*(z,\theta)})d\theta dz = \\ &= \frac{-1}{T+1}\int p_T(\theta)\int p_T^*(z)\log(\frac{p_{\zeta,T}(z|\theta)}{p_T^*(z)})dzd\theta = \\ &= \boldsymbol{E}_{\text{ouv}}(\frac{-1}{T+1}\boldsymbol{E}_{p_T^*}\log(\frac{p_{\zeta,T}(.|\theta)}{p_T^*})) \end{aligned} \tag{4.92}$$

By this measure, we are taking the mean of the Kullback-Leibler divergence from model number T to process number T over all possible values of the process θ that generated the coefficients of the model, weighting these values by our OUV prior distribution. An important aspect of this version of the Kullback-Leibler divergence is that it is non-negative and invariant for (bijective) reparametrizations of the sample- and/or parameter space. (See S.Kullback, 1959, chapter two) [9] [10].
Hence, in order to obtain optimal fitting of a model from the class to the given sequence of processes, one should minimize the asymptotic Kullback-Leibler divergence (w.r.t. ζ) over both the sample- and parameter space,

$$D(p_\zeta, p^*, Sam \times Par) = \lim_{T\uparrow\infty} \boldsymbol{E}_{\text{ouv}} D_T(p_\zeta, p^*|\theta) \text{ where} \tag{4.93}$$

$$D_T(p_\zeta, p^*|\theta) = \frac{-1}{T+1}\boldsymbol{E}_{p_T^*}\log(\frac{p_{\zeta,T}(.|\theta)}{p_T^*}), \tag{4.94}$$

the conditional Kullback-Leibler divergence, a $\mathcal{B}_1$-measurable function (we assume that the limit exists). We shall base the criterion that we want to minimize on this asymptotic Kullback Leibler divergence.
In equation (4.93), we can split the expectation over the parameter space in two terms:

$$\begin{aligned} D(p_\zeta, p^*, Sam \times Par) &= D_1(\zeta) + D_2(\zeta), \text{ where} \\ D_1(\zeta) &= \lim_{T\uparrow\infty} \boldsymbol{E}_{\text{ouv}} D_T(p_\zeta, p^*|\theta)1_{\{\nu_\zeta > 1\}} \text{ and} \qquad (4.95) \\ D_2(\zeta) &= \lim_{T\uparrow\infty} \boldsymbol{E}_{\text{ouv}} D_T(p_\zeta, p^*|\theta)1_{\{\nu_\zeta \leq 1\}}. \qquad (4.96) \end{aligned}$$

We discuss the second term, $D_2(\zeta)$.
On the event $\{\nu_\zeta \leq 1\}$, BIBO stability, or the existence of a conditional spectral density,

[9] By reparametrizing one can also include the transition variance parameters τ_i in the set of hyper-parameters, we propose to estimate by this method.

[10] Non-negativity and invariance are shared by other types of divergences, based on ratios of probability densities too (see e.g. Amari, 1990, p.87), but for *time-invariant* stationary autoregressive processes, Amari (1987) has given geometrical arguments, showing that the Kullback-Leibler divergence is a very natural one.

is not guaranteed. If we work with a model class satisfying the assumptions of part 4 of theorem 4.13, then stability and spectral density are even excluded. Our assumption about the given sequence of processes was, however, that it has a density, hence, that it is BIBO stable. Accordingly, we may expect that the models, corresponding to a θ in the event $\{\nu_\zeta \leq 1\}$, have large divergence to the given sequence, much larger than the models corresponding to the event $\{\nu_\zeta > 1\}$. In other words, we have an a priori dislike in fitting an unstable model to the given sequence. We can lay more stress on this dislike in our model by one type of specification of the model we did not completely make until now: we did not completely define what the instantaneous polynomial $q_{t+1,T}$ should be *after* the running time ν_T. That is, we did not define the $h_{t+1,T}$ in equation (4.46). As an alternative to assumption 6 of theorem 4.13, we can define: let $h_{t+1,T} = h$, h $\mathcal{B}_0$-measurable, $h : \Omega \to \mathbb{R}^{n+1} \setminus U_n$. It is easy to adapt the definitions of General Model etc. in order to enable us to make this kind of assumption. If we do so, part 4 of the Theorem 4.13 will remain valid. In other words: as soon as the coefficients of the model menace to run out of U_n, we stop the coefficient process of the model and make the coefficient (or instantaneous associated polynomial) a constant, unstable polynomial. By choosing h far or close to U_n, we can manipulate the divergence of a model with corresponding coefficient process to the given sequence. In this way, we can give prior weights to our feeling that the given sequence of processes is really BIBO stable. But to make things less complicated, we choose for another option with the same idea. In the choice of the criterion we replace term $D_2(\zeta)$ simply by:

$$\tilde{D}_2(\zeta) = \mathcal{M}P(\{\nu_\zeta \leq 1\}) \qquad (\mathcal{M} \text{ constant, not depending on } \theta \text{ nor } \zeta). \tag{4.97}$$

Here, the number $\mathcal{M}$ should be large.
Now, we discuss the first term, $D_1(\zeta)$.
The models, corresponding to the event $\{\nu_\zeta > 1\}$, admit the spectral density, hence on this event, we have point-wise convergence in the parameter space Par of $D_T(p_\zeta, p^*|\theta)$ to the expression (4.86), derived by Dahlhaus. We have to derive conditions, such that $D_T(p_\zeta, p^*|\theta)$ is majorized by an integrable function of θ. Then, we can apply the majorized convergence theorem of Lebesgue, and rewrite term D_1 as

$$D_1(\zeta) = \boldsymbol{E}_{\text{ouv}}\left(\frac{1}{2}\int_0^1 \int_{\mathbf{I}} \left\{\log\left(\frac{f_\zeta(u,w)}{f^*(u,w|\theta)}\right) + \frac{f^*(u,w|\theta)}{f_\zeta(u,w)} - 1\right\} dm(w) du 1_{\{\nu_\zeta > 1\}}\right). \tag{4.98}$$

Dropping the terms which do not depend on ζ, we now obtain the first criterion that we propose to minimize w.r.t. ζ:

$$\mathcal{J}(\zeta) := \boldsymbol{E}_{\text{ouv}}\left(\frac{1}{2}\int_0^1 \int_{\mathbf{I}} \left\{\log(f_\zeta(u,w|\theta)) + \frac{f^*(u,w)}{f_\zeta(u,w|\theta)}\right\} dm(w) du 1_{\{\nu_\zeta > 1\}}\right) + \mathcal{M}P(\nu_\zeta \leq 1). \tag{4.99}$$

We follow the idea of Dahlhaus to estimate the spectral density of the given sequence of processes by some kind of periodogram estimator $I_N(u,w)$, and propose the following actual criterion to be minimized w.r.t. ζ:

$$\mathcal{J}_T(\zeta) := \boldsymbol{E}_{\text{ouv}}\left(\frac{1}{2}\frac{1}{S}\sum_{j=1}^{S} \int_{\mathbf{I}} \left\{\log(f_\zeta(u_j,w|\theta) + \frac{I_N(u_j,w)}{f_\zeta(u_j,w|\theta)}\right\} dm(w) 1_{\{\nu_\zeta > 1\}}\right) + \mathcal{M}P(\nu_\zeta \leq 1). \quad (u_j = \frac{j}{S}) \tag{4.100}$$

We are not going to do any asymptotic analysis of the variables ζ_0 resp. $\hat{\zeta}_T$, minimizing the two criteria $\mathcal{J}(\zeta)$ and $\mathcal{J}_T(\zeta)$ respectively. That is beyond the scope of this thesis. We only show the link between minimizing the second criterion, and the ad hoc method we used in experiments (see chapter six). In this method, we simply excluded values of ζ that yielded coefficient estimates running out of U_n, and under this restriction, maximized the likelihood with respect to ζ. We computed this (approximate) likelihood by Prediction Error Decomposition, included in an Extended Kalman Filtering algorithm.
Also in showing this link, we restrict ourselves. We make a very special choice of spectral density estimator and very special choices of N and S. The justification of this restriction is that we only want to illustrate ideas.
A simple periodogram-type of the spectrographic estimator (without data-taper h as in equation (4.88)) is

$$I_N(u,w) = \frac{1}{N}|\sum_{k=0}^{N-1} \bar{w}^k y_{[uT]-\frac{N}{2}+k+1,T}|^2. \tag{4.101}$$

The spectral density $f_\zeta(u,w|\theta)$ is, according to proposition 4.6 :

$$f_\zeta(u,w|\theta) = \frac{\sigma(u,\zeta)^2}{|w^n - a_1(u,\zeta)w^{n-1} - \cdots - a_n(u,\zeta)|^2} \tag{4.102}$$

where $\gamma(u,\zeta) = \frac{1}{\sigma(u,\zeta)}(w^n - a_1(u,\zeta)w^{n-1} - \cdots - a_n(u,\zeta))$. With this spectral density-estimator and expression for f_ζ, the actual criterion $\mathcal{J}_T(\zeta, y_{\cdot,T})$ becomes:

$$\begin{aligned}
\mathcal{J}_T(\zeta, y_{\cdot,T}) &= \frac{1}{2(S-1)} \sum_{i=1}^{S-1} \boldsymbol{E}_{\text{ouv}}\{(\log(\sigma^2(u_i,\zeta)) + \\
&\quad + \frac{1}{\sigma^2(u_i,\zeta)} t(u_i))1_{\{\nu_\zeta > 1\}}\} + P(\{\nu_\zeta \leq 1\})\mathcal{M} \quad (u_i = \frac{i}{S}), \quad \text{where} \qquad (4.103)\\
t(u) &= \frac{1}{N}\int_{\mathbf{I}} |\sum_{k=0}^{N-1} \bar{w}^k(1 - a_1(u,\zeta)\bar{w} + \\
&\quad - \cdots - a_n(u,\zeta)\bar{w}^n) y_{[uT]-\frac{N}{2}+k+1,T}|^2 dm(w) = \\
&= \frac{1}{N}\int_{\mathbf{I}} |\sum_{k=n}^{N-1} (\tilde{y}_k - \tilde{a}_1\tilde{y}_{k-1} - \cdots \tilde{a}_n\tilde{y}_{k-n})\bar{w}^k|^2 dm(w) + \epsilon(u,y,N).
\end{aligned}$$

Here, $\tilde{y}_k = y_{[uT]-\frac{N}{2}+k+1,T}$, $\tilde{a}_i = a_i(u,\zeta)$, and $\epsilon(u,y,N)$ is given by:

$$\epsilon(u,y,N) = \frac{1}{N}\int_{\mathbf{I}} |\sum_{s=0}^{n-1} (\tilde{a}_{1+s}\tilde{y}_{N-1} + \cdots + \tilde{a}_n\tilde{y}_{N+s-n})\bar{w}^{N+s}|^2 dm(w). \tag{4.104}$$

Now, we can rewrite $t(u)$ as follows:

$$\begin{aligned}
t(u) &= \frac{1}{N}\sum_{k=n}^{N-1} (\tilde{y}_k - \tilde{a}_1\tilde{y}_{k-1} - \cdots \tilde{a}_n\tilde{y}_{k-n})^2 + \epsilon(u,y,N) = \\
&= \frac{1}{N}\sum_{k=[uT]-\frac{N}{2}+n+1}^{[uT]+\frac{N}{2}+n} (y_{k,T} - a_{1,k,T}y_{k-1,T} - \cdots - a_{n,k,T}y_{k-n,T})^2 + \\
&\quad + \tilde{\epsilon}(u,y,N). \qquad (4.105)
\end{aligned}$$

If $NS = T + 1$, we find that the actual criterion can be rewritten as:

$$\begin{aligned} & \mathcal{J}_T(\zeta, y_{0,T}, \cdots, y_{T,T}) = \tilde{\mathcal{J}}_T + P(\{\nu_\zeta \leq 1\})\mathcal{M} + \eta(y, N, T) \quad , \text{where} \\ \tilde{\mathcal{J}}_T \quad = \quad & \boldsymbol{E}_{\text{ouv}}\{ \, 1_{\{\nu_{\zeta,T} > T, \nu_\zeta > 1\}} \frac{1}{2(T+1)} \times \\ & \times \sum_{k=0}^{T} (\log(\sigma^2_{k,T}) + \frac{(y_{k,T} - a_{1,k,T} y_{k-1,T} - \cdots - a_{n,k,T} y_{k-n,T})^2}{\sigma^2_{k,T}}).\} \\ & . \end{aligned} \tag{4.106}$$

Statement For a suitable choice of N, e.g. $N = T^{\frac{1}{4}}$, a suitable choice of the manifold of the sequence of models (e.g. compact subset with non-empty interior of U_n), and a uniformly bounded sequence of processes $(y_{t,T})_{0 \leq t \leq T}$, $T \in \mathbb{N}$, we have:

$$\lim_{T \uparrow \infty} \eta(y, N, T) \quad = \quad 0. \tag{4.107}$$

Proof: We only give elements of the proof of this statement. It is immediate from equation (4.104) that

$$\epsilon(u, y, N) = O(\frac{1}{N}) \quad (N \to \infty).$$

By the compactness of the manifold in U_n the concept of "bounded and equislowly varying coefficients" can be defined equivalently in terms of the metric of the model and in terms of the Euclidean metric. We even have:

$$\boldsymbol{E}_{\text{ouv}} \boldsymbol{d}^2(\begin{pmatrix} a_{k+1,T} \\ \sigma^2_{k+1,T} \end{pmatrix}, \begin{pmatrix} a_{k,T} \\ \sigma^2_{k,T}) \end{pmatrix}) 1_{\{\nu_{\zeta,T} > T, \nu_\zeta > 1\}} = \frac{1}{T^2} \boldsymbol{E}_{\text{ouv}} \theta^2(\frac{k}{T}) 1_{\{,\nu_{\zeta,T} > T, \nu_\zeta > 1\}} \leq \frac{C}{T^2}$$

for some $C \in \mathbb{R}_+$, where $\boldsymbol{d}$ is the metric of the model, and a similar inequality holds w.r.t. the Euclidean metric. Because the $y_{t,T}$ are uniformly bounded, and the coefficients $(a_{[uT],T}, \sigma^2_{[uT],T})$ uniformly converge to $(a(u,\zeta), \sigma^2(u,\zeta))$ in mean square, according to proposition 4.10, it is easy to deduce that

$$\lim_{T \uparrow \infty} \tilde{\epsilon}(u, y, N) = 0$$

and to finish the proof. □

For large $\mathcal{M}$ and large T, we can approximate the minimizing ζ of $\mathcal{J}_T$ by minimizing the function $\tilde{\mathcal{J}}_T(\zeta)$ over a subset $\mathcal{Z}_0$ of the parameter space, where $P(\{\nu_\zeta > 1\})$ is large (and almost constant).
Minimizing expression $\tilde{\mathcal{J}}_T$ (4.106) over $\mathcal{Z}_0$ is the same as *maximizing the likelihood* $p_{\zeta,T}(y_{0,T}, \cdots, y_{T,T})$ (conditional on a running time larger than T) over $\mathcal{Z}_0$, as will be shown in Remark 6.10.
Above, we have put the ordinary maximum likelihood method by prediction error decomposition for estimating the "extra structure", in perspective. We have equated it to a form of criterion $\mathcal{J}_T$ of equation (4.99), in fact the most rude form because we used the most simple periodogram-type of spectrographic estimator. This suggests that improvements can be made: introducing data-tapers in the spectral density estimator; choosing

other combinations of N and S; taking more account of the event $\{\nu_\zeta \leq 1\}$. This section had a very heuristic character; almost everything in this direction has to be investigated yet. Despite this, we think that a merge between the well-analyzed methods of Dahlhaus and the stochastic coefficient modelling methods is possible, yielding the possibility of parsimonious models with higher precision.

In this section we did not discuss how to choose the prior OUV distribution. The determination problem of transition matrix hyperparameters β_i and transition variance hyperparameters τ_i is common to all models, where time-varying parameters are assumed to follow (smoothly integrated) random walks or generalizations of these. We did not discuss that problem here. We shall discuss this subject in chapter six. What we elucidated in this chapter was just isolating the remaining problems (the determination of the other hyperparameters) from this crucial problem. In this respect, we did a similar thing as in chapter three, section 3.8: there we isolated a linear transition equation with parameters β_i and τ_i, leaving a non-linear measurement equation.

Chapter 5

Geometry on U_n

This chapter deals with geometrical structures on the set U_n of stable polynomials of degree n with positive highest coefficient, or, equivalently, the set of stationary AR(n) models. The theories, underlying this section, have been gathered in chapter two: the section on the theory of shift invariant inner products; the section on the asymptotic Fisher matrix for a univariate constant AR(n) coefficient process; the introductory section on differential geometry, and the subsection on submanifolds and curvature.
U_n represents the class of stationary AR(n) processes. Different parametrizations of (parts of) U_n have been studied in statistics. Usually, the purpose in these studies is to obtain certain desired properties of the asymptotic distribution of ML estimators of these parameters for AR(n) processes.
Such a study has been done in general for models in the class of exponential families. A Gaussian stationary AR(n)-process $y_0, \cdots, y_T$ can be regarded as a member of a curved $(\frac{(n+1)(n+2)}{2}, n+1)$ exponential family, as is shown by Arato in (1961).
For one-dimensional curved exponential families, different types of parametrizations were studied by Hougaard (1982). He gave some rules to obtain parametrizations, such that the corresponding MLE satisfies various asymptotic requirements, such as

- the MLE has zero asymptotic skewness;

- the MLE has a stabilized variance, i.e. a variance that does not depend on the parameter value.

By Kass (1984) it was pointed out that the rules of Hougaard could be geometrically interpreted. These rules were in fact flatness criteria w.r.t. different connections of the Riemannian geometry given by the Fisher metric (see also Amari, 1990, pp.150-152). Parametrizations, yielding zero asymptotic skewness, are possible if and only if the corresponding (sub)manifold is flat in the $-\frac{1}{3}$-connection w.r.t. the Fisher metric. Covariance stabilizing parametrizations are possible if and only if the corresponding (sub)manifold has zero curvature in the 0-connection w.r.t. the Fisher metric. The 0-connection is the usual Levi-Civita-connection, the only type of connection that we use in this thesis.
The importance of covariance stabilizing parametrizations has been pointed out by Vaeth (1985). Unlike the likelihood ratio test statistic, the Wald test depends on the used parametrization. For one-parameter exponential families, Vaeth showed that the Wald

test may-be ill-behaved unless we use covariance stabilizing parametrizations.
For AR(1) models, Ravishankar, Melnick and Tsai (1990) computed these types of parametrizations. The parametrization that reduces the skewness of the distribution of the MLE asymptotically to zero, is the one where the AR(1)-coefficient $a \in (-1,1)$ in the equation $y_t = ay_{t-1} + \sigma\epsilon_t$ is replaced by:

$$a = \tanh(\kappa) := \frac{e^{\kappa} - e^{-\kappa}}{e^{\kappa} + e^{-\kappa}} \quad (\kappa \in \mathbb{R}). \tag{5.1}$$

The covariance stabilizing parametrization that makes the asymptotic variance of the MLE constant on the parameter space, is the one given by:

$$a = \sin(\kappa) \qquad (\kappa \in (-\frac{\pi}{2}, \frac{\pi}{2})). \tag{5.2}$$

Both parametrizations were also used in a different context by Gray and Markel (1976).

A main objective of our study of geometrical structures of U_n is to design General Models that have a linear transition equation with respect to some parametrization of a submanifold of U_n. Such models will be named *zero-curvature* models. In chapter three we gave sufficient and necessary conditions for a General Model to have a linear transition equation. We shall study a few interesting parametrizations, for different metrics, with respect to which the transition equation can be linear. We are especially interested in their analytical formulas.
We shall *first* investigate the geometry given by the *Fisher metric* and the Levi-Civita-connection. From one point of view, one can see our study as a search for *covariance stabilizing* parametrizations. For U_n, it will be shown that for dimensions higher than one there do not exist covariance stabilizing parametrizations of the whole manifold, i.e. parametrizations, such that the asymptotic variance of the MLE does not depend on the parameter values. But there do exist parametrizations that stabilize the asymptotic MLE covariance of *one* parameter with respect to itself and the others. We mean parametrizations, such that the MLE of *one* parameter has asymptotic variance constant on the manifold and, moreover, is asymptotically uncorrelated with the MLEs of the other parameters. These parametrizations will be called *pencils of geodesics.*

Secondly, we introduce a complete metric which, in the AR(1) case, yields the asymptotic skewness reducing parametrization. Furthermore, we precisely define what we mean by *Euclidean* metric.

As already was indicated, throughout this section, we work with *parametrizations* $\psi : U \subset \mathbb{R}^{n+1} \to U_n$, instead of with their inverses, the *charts* $x = \psi^{-1} : U_n \to U$, just as was done in the introductory section on differential geometry in chapter two.

5.1 The asymptotic Fisher metric on U_n

In order to define a Riemannian structure on the set U_n, we first have to turn U_n into a differential manifold. This is simple, because U_n is an open and connected subset of

the set Π_n of all polynomials of degree n. Π_n will be identified with $\mathbf{R}^{n+1}$ by the trivial map: $(b_0, \ldots, b_n) \mapsto b_0 x^n + \ldots + b_n$. The tangent spaces in every point of U_n are also identified in the usual way with $\mathbf{R}^{n+1}$, and thus with Π_n, so that the tangent vector $\frac{\partial}{\partial b_i}$ is identified with x^{n-i}.

Now we introduce a distance on U_n. As we have linked the elements of U_n to the Gaussian stationary AR(n)systems, it seems evident that a good concept of distance between these systems will give a relevant geometrical structure on U_n.
A good notion of distance between Gaussian stationary AR(n) systems can be based on the *asymptotic Fisher information.*

Consider a Gaussian stationary AR(n) process:

$$z_t = a_1 z_{t-1} + \ldots + a_n z_{t-n} + \sigma e_t \tag{5.3}$$

where $(e_t)_{t \in \mathbf{Z}}$ i.i.d.$\mathcal{N}(0,1)$.
Its associated polynomial is $L_n = x^n - a_1 x^{n-1} - \cdots - a_n$; we have seen that $L_n \in U_n$. Let $\Theta = (a_1, \ldots, a_n, \sigma^2)$ be its *natural parameter*, and let η: $\eta_i := E z_t z_{t-i}$ if $i \in \mathbf{Z}_+$ be its *covariance function* . The map

$$\Theta : U_n' = \{(a_1, \ldots, a_n, \sigma^2) | \frac{1}{\sigma}(x^n - a_1 x^{n-1} - \ldots - a_n) \in U_n\} \to U_n$$

defined by:

$$(a_1, \ldots, a_n, \sigma^2) \mapsto L = \frac{1}{\sigma}(x^n - a_1 x^{n-1} - \ldots - a_n)$$

is C^∞-differentiable, and so is its inverse, hence U_n', Θ is a parametrization of U_n. The asymptotic Fisher matrix of the Gaussian stationary AR(n) process (5.3) w.r.t. the natural parameters $(a_1, \ldots, a_n, \sigma^2)$ is equal to:

$$H_\Theta^n = \begin{pmatrix} \frac{\eta_0}{\sigma^2} & \frac{\eta_1}{\sigma^2} & \ldots & \frac{\eta_{n-1}}{\sigma^2} & 0 \\ \frac{\eta_1}{\sigma^2} & \frac{\eta_0}{\sigma^2} & \ldots & \frac{\eta_{n-2}}{\sigma^2} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\eta_{n-1}}{\sigma^2} & \frac{\eta_{n-2}}{\sigma^2} & \ldots & \frac{\eta_0}{\sigma^2} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2\sigma^4} \end{pmatrix}, \tag{5.4}$$

(see chapter two). H_Θ^n is positive definite, and equation (2.22) shows that the matrix elements $\frac{\eta_i}{\sigma^2}$ depend on the natural parameters $(a_1, \ldots, a_n, \sigma^2)$ in a C^∞-differentiable way. As U_n', Θ is a parametrization of U_n, H_Θ^n defines a Riemannian structure on U_n.

The Riemannian structure of ARMA(n,q) processes, based on the asymptotic Fisher information matrix, has been studied before. The most important contribution has been made by Amari (1987). He has shown that in the set of AR(n) models, it makes much sense to use the so called 1- and -1-connections instead of the Levi Civita connection, because U_n is flat with respect to these connections. As a consequence, instead of the Riemannian distance, Amari has to work with the weaker concept of the Kullback-Leibler divergence. We shall refer to his article later. Another article on this subject was

written by Ravishanker, Melnick and Tsai(1990). We already referred to their article above. In addition to what has been said there, these authors computed formulas for Fisher matrix and Christoffel symbols that determine the geometrical connections for ARMA(n,q) models. We shall study the Riemannian geometry of U_n in a way that differs from both Amari and Ravishankar, Melnick and Tsai:

- By characterizing the Fisher metric on U_n in a coordinate free way;
- By studying the Riemannian distance and not any other divergence; consequently, we work with the Levi Civita connection and not with other connections;
- By studying the pencils of geodesics perpendicular to two types of hypersurfaces: the hypersurfaces of polynomials in U_n with constant highest coefficient, and the hypersurfaces of polynomials in U_n with constant last Schur parameter;
- By studying geometrical properties like geodesical completeness and sectional curvature;
- By giving all the geodesics in U_2.

5.1.1 The pencil of noise geodesics and characterization of the Fisher metric

First, we study the hypersurfaces of polynomials with constant highest coefficient, i.e. the polynomials corresponding to a fixed noise-level. We shall show that these are totally geodesic. Simultaneously, we give a coordinate free characterization of the Fisher metric.

Definition. AR_n is the hypersurface in U_n of stable polynomials of degree n which are monic, i.e. correspond to a noise level equal to one.

Alternative to the natural parametrization.
Instead of working with

$$\Theta : U_n' \to U_n; \quad (a_1, \cdots, a_n, \sigma^2) \overset{\Theta}{\mapsto} L = \frac{1}{\sigma}(x^n - a_1 x^{n-1} - \cdots - a_n)$$

we can put $r = \log \sigma$. Then we obtain another parametrization χ of U_n, defined by:

$$\chi : U_n'' = \{(a_1, \ldots, a_n, r) | e^{-r}(x^n - a_1 x^{n-1} \ldots - a_n) \in Un\} \to$$

$$\to U_n;$$

$$(a_1, \ldots, a_n, r) \mapsto e^{-r}(x^n - a_1 x^{n-1} \ldots - a_n).$$

The coordinate vector field $\frac{\partial}{\partial r} = \frac{\partial \chi}{\partial r}$ will be called the *log-noise vector field.*

Proposition 5.1

1. **(coordinate free characterization of the asymptotic Fisher metric).** The inner product, defined on the tangent space Π_n in the point $L_n \in U_n$ by the asymptotic Fisher matrix, is related to the Schur inner product on Π_n, induced by L_n in the following way:
   (a) both inner products coincide on Π_{n-1};
   (b) both inner products L_n are perpendicular to Π_{n-1};
   (c) according to the Schur inner product $\langle L_n, L_n \rangle_{Schur} = 1$, and according to the Fisher inner product $\langle L_n, L_n \rangle_{Fisher} = 2$.

2. $\frac{1}{\sigma} AR_n$ is isometric isomorphic with AR_n by the diffeomorfism : $AR_n \to \frac{1}{\sigma} AR_n$; $L_n \mapsto \frac{1}{c} L_n$;

3. The log-noise vector field $\frac{\partial}{\partial r}$ is perpendicular to any submanifold $\frac{1}{\sigma} AR_n$ of U_n, and satisfies:
$$\nabla_X \frac{\partial}{\partial r} = 0 \tag{5.5}$$
for any vector field X on U_n. AR_n is a totally geodesic submanifold of U_n;

4. For any $L_n \in U_n$, the curve $r \mapsto e^{-r} L_n$ ($r \in \mathbb{R}$) is a geodesic.

Remark 5.2 *The geodesics $r \mapsto e^{-r} L_n$ will be called the noise geodesics.* □

We like to recall here that the main facts about Schur inner products were summarized in the section about the theory of shift invariant inner products in chapter two.

Proof:

1. With respect to the newly introduced parametrization χ, the asymptotic Fisher metric becomes (use equation (2.4)):
$$H_\chi^n = \begin{pmatrix} \frac{\eta_0}{e^{2r}} & \frac{\eta_1}{e^{2r}} & \dots & \frac{\eta_{n-1}}{e^{2r}} & 0 \\ \frac{\eta_1}{e^{2r}} & \frac{\eta_0}{e^{2r}} & \dots & \frac{\eta_{n-2}}{e^{2r}} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\eta_{n-1}}{e^{2r}} & \frac{\eta_{n-2}}{e^{2r}} & \dots & \frac{\eta_0}{e^{2r}} & 0 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix}. \tag{5.6}$$
   Furthermore, in the point L_n the tangent vector $\frac{\partial}{\partial r} = -L_n$, hence $\langle L_n, L_n \rangle = 2$. The tangent spaces of U_n are spanned by the vectors $\frac{\partial}{\partial a_i}$ and $\frac{\partial}{\partial r}$. Now, we have:
   $x^i = \frac{\partial}{\partial b_{n-i}} = -e^r \frac{\partial}{\partial a_{n-i}}$, if $i < n$, hence $\langle x^i, x^j \rangle = \eta_{i-j}$ for all $i, j < n$;
   $-L_n = \frac{\partial}{\partial r}$ is perpendicular to $\frac{\partial}{\partial a_{n-i}}$ for all $i < n$, hence L_n is perpendicular to Π_{n-1}.
   These two facts are also true for the Schur inner product, induced by L_n (see chapter two, equation (2.29) and equation (2.22)). We have seen above that $\langle L_n, L_n \rangle_{Fisher} = 2$, and in chapter two that $\langle L_n, L_n \rangle_{Schur} = 1$.

2. Note from equation (2.22, chapter two) that the matrix elements $\frac{\eta_{ij}}{\sigma^2}$ of H^n_χ do not depend on σ, only on $(a_1,\ldots,a_n)$. Thus, the left upper part of the matrix H^n_χ defines isometric Riemannian structures on all the hypersurfaces $\frac{1}{\sigma}AR_n$, $\sigma \in \mathbb{R}^+$.

3. The tangent spaces of $\frac{1}{\sigma}AR_n$ are spanned by the vectors $\frac{\partial}{\partial a_i}$, and from H^n_χ we see that $\frac{\partial}{\partial r}$ is perpendicular to these vectors. We have already noted that the elements of the matrix H^n_χ do not depend on σ, thus also not on r. If we use the properties of the Levi-Civita-connection, as mentioned in chapter two, then equation (5.5) is immediate. Hence, the second fundamental form of AR_n w.r.t. U_n is zero and, consequently, geodesics in AR_n are also geodesics in U_n (see chapter two, subsection about submanifolds)

4. The tangent vector $\frac{\partial}{\partial r}$ of the curve $r \mapsto e^{-r}L_n$ satisfies $\nabla_{\frac{\partial}{\partial r}}\frac{\partial}{\partial r} = 0$. □

By the third part of proposition 5.1 it is clear that any parametrization $\tilde{\psi} : \left\{ \begin{array}{ccc} V & \to & AR_n \\ \kappa & \mapsto & \tilde{\psi}(\kappa) \end{array} \right\}$ can be extended to a parametrization $\psi : \left\{ \begin{array}{ccc} V \times \mathbb{R} & \to & U_n \\ (\kappa, r) & \mapsto & (\tilde{\psi}(\kappa), r) \end{array} \right\}$, such that the asymptotic Fisher matrix with respect to this parametrization has the form

$$H_\psi(\kappa, r) = \begin{pmatrix} H_{\tilde{\psi}}(\kappa) & 0 \\ 0 & 2 \end{pmatrix}.$$

Hence, the MLE of the parameters κ_i of a stationary AR(n) process has asymptotic variance that does not depend on the log-noise-level r, and that is uncorrelated with the MLE of this log-noise-level. This fact is reflected in the possibility to concentrate the noise-level out of the likelihood for such a process; the noise-level can be regarded as just a scaling factor. Parametrizing the noise-level by its logarithm stabilizes the variance of its MLE.

5.1.2 The pencil of Schur geodesics and sectional curvature

In propositon 5.1 item 1, we found the pencil of geodesics, described by the noise level (the noise geodesics). For further study of the geometry of U_n, it suffices to study AR_n, because we have seen in proposition 5.1 item 2 that AR_n is isometrically isomorphic to $\frac{1}{\sigma}AR_n$ for all $\sigma \in \mathbb{R}^+$, and AR_n is totally geodesic. We now present another pencil of geodesics, directed on AR_n, perpendicular to the hypersurfaces of all polynomials in AR_n with constant last Schur parameter. For that purpose, we shall use the theory of Schur parameters, as stated in chapter two, in particular the theorem of the last Schur parameter.

Definition. AR_{n-1} will be inbedded isometrically in AR_n by:$L_{n-1}(x) \mapsto xL_{n-1}(x)$. We denote the set of polynomials in AR_n with the same last Schur parameter $\lambda = c$ by: $AR_n^{\lambda=c}$. Then, $AR_n^{\lambda=0} = AR_{n-1}$.

Proposition 5.3

1. For any parametrization $\rho : V \to AR_{n-1}, (d_1, \ldots, d_{n-1}) \mapsto xL_{n-1}(x)$ of AR_{n-1}, we obtain a new parametrization of AR_n by defining: $\tilde{\rho} : V \times (-\frac{\pi}{2}, \frac{\pi}{2}) \to AR_n; (d_1, \ldots, d_{n-1}, \alpha) \mapsto L_n(x) = xL_{n-1}(x) + \sin(\alpha)E_{n-1}(x)$, where $E_{n-1}(x) = P_n(xL_{n-1}(x))$ is the sip transform of $xL_{n-1}(x)$;

2. For any $L_n = xL_{n-1} + \sin(\alpha)E_{n-1} \in AR_n$, the curve $\alpha \mapsto xL_{n-1} + \sin(\alpha)E_{n-1}$ $(\alpha \in (-\frac{\pi}{2}, \frac{\pi}{2}))$ is a geodesic of maximal length on AR_n. AR_n and U_n are not geodesically complete; the distances on AR_n are bounded from above by $2n\pi$;

3. The map $f : AR_{n-1} \to AR_n^{\lambda=\sin(\alpha)}; xL_{n-1} \mapsto L_n = xL_{n-1} + \sin(\alpha)E_{n-1}$ is not isometric, but satisfies:

$$\langle \frac{df \circ \gamma}{dt}_{|0}, \frac{df \circ \mu}{dt}_{|0} \rangle_{|L_n} = \frac{1+\sin^2(\alpha)}{\cos^2(\alpha)} \langle \frac{d\gamma}{dt}_{|0}, \frac{d\mu}{dt}_{|0} \rangle_{xL_{n-1}} + \frac{2\sin(\alpha)}{\cos^2(\alpha)} \langle P_n(\frac{d\gamma}{dt}_{|0}), \frac{d\mu}{dt}_{|0} \rangle_{|xL_{n-1}},$$

i.e.

$$\langle \frac{\partial}{\partial d_i}, \frac{\partial}{\partial d_j} \rangle_{|L_n} = \frac{1+\sin^2(\alpha)}{\cos^2(\alpha)} \langle \frac{\partial}{\partial d_i}, \frac{\partial}{\partial d_j} \rangle_{|xL_{n-1}} + \frac{2\sin(\alpha)}{\cos^2(\alpha)} \langle P_n(\frac{\partial}{\partial d_i}), \frac{\partial}{\partial d_j} \rangle_{|xL_{n-1}} \quad (5.7)$$

4. **(Weingarten's map for $AR_n^{\lambda=\sin(\alpha)}$)**
The vector field $\frac{\partial}{\partial \alpha}$ is perpendicular to any submanifold $AR_n^{\lambda=\sin(\alpha)}$ of AR_n and satisfies :

$$\nabla_X \frac{\partial}{\partial \alpha} = \frac{1}{\cos(\alpha)} P_n(X) \quad (5.8)$$

for any vector field X on $AR_n^{\lambda=\sin(\alpha)}$. A geodesic on the submanifold AR_{n-1} is in general *not* a geodesic in the manifold AR_n. AR_{n-1} is not a totally geodesic submanifold of AR_n.

Remark 5.4 *The curves $\alpha \mapsto xL_{n-1}(x) + \sin(\alpha)E_{n-1}(x)$ will be called the Schur geodesics.* □

Proof:

1. That $\tilde{\rho}$ is bijective follows directly from the theorem of the last Schur parameter, which has been stated in chapter two. The C^∞-differentiability of $\tilde{\rho}$ and its inverse is easily verified.

2. If $L_n = \tilde{\rho}(d_1, \ldots, d_{n-1}, \alpha) = x^n + \ldots + b_n$, then $b_n = L_n(0) = \sin(\alpha)$. Hence, $\frac{\partial}{\partial d_i} b_n = 0$, thus $\frac{\partial}{\partial d_i}$ is a linear combination of the vectors $\frac{\partial}{\partial b_j} = x^{n-j}$ with $1 \leq j < n$, i.e. $\frac{\partial}{\partial d_i} \in x\Pi_{n-2}$. The tangent space in L_n to $AR_n^{\lambda=\sin(\alpha)}$ is, consequently, $x\Pi_{n-2}$ whereas the tangent space in L_n to AR_n is Π_{n-1}, as we have seen before. The Fisher inner product on the tangent space in L_n to AR_n coincides with the Schur inner product on Π_n, induced by L_n. We know from the theorem of the last Schur parameter that the restriction of this inner product to Π_{n-1} is induced by $\sqrt{(1-\lambda^2)}L_{n-1}=\cos(\alpha)L_{n-1}$. Hence, $E_{n-1} = P_{n-1}(L_{n-1})$ is a multiple of the "zero

evaluator" in Π_{n-1}, from which it follows that E_{n-1} is perpendicular to $x\Pi_{n-2}$. We have also that $\langle \cos(\alpha)E_{n-1}, \cos(\alpha)E_{n-1}\rangle = \langle cos(\alpha)L_{n-1}, \cos(\alpha)L_{n-1}\rangle = 1$. However, $\frac{\partial}{\partial\alpha} = \cos(\alpha)E_{n-1}$, hence the Fisher metric on the parameters $d_1, \ldots, d_{n-1}, \alpha$ has the form:

$$H_{\tilde{\rho}}^{AR_n} = \begin{pmatrix} h(d_1, \ldots, d_n, \alpha) & 0 \\ 0 & 1 \end{pmatrix} \tag{5.9}$$

According to chapter two, this shows that

$$\alpha \mapsto \tilde{\rho}(d_1, \ldots, d_n, \alpha),\ \alpha \in (-\frac{\pi}{2}, \frac{\pi}{2}),\ (d_1, \ldots, d_n) \in V$$

defines a pencil of geodesics. As $\tilde{\rho}$ is surjective these geodesics can't be extended; they are maximal. Accordingly, AR_n (and also U_n) is geodesically not complete. The length of the geodesics of this pencil (which are described by the last Schur parameter) is π. Applying the theorem of the last Schur parameter n times yields the unique set of n Schur parameters of the polynomial $L_n \in AR_n$, and the stable polynomials $xL_{n-1}, \ldots, x^{n-1}L_1, x^nL_0 = x^n$. We shall describe a curve from the polynomial x^n to L_n. First, go from x^n to $x^{n-}L_1$ following the geodesic on the submanifold AR_1, described by the first Schur parameter. Next, go from $x^{n-1}L_1$ to $x^{n-2}L_2$ following the geodesic on the submanifold AR_2, described by the second Schur parameter and so on. All these curves have length less than π on their submanifolds, but as these submanifolds are isometrically inbedded in AR_n; their length remains the same in AR_n. So, the total length of the curve from x^n to L_n is less than $n\pi$, hence all distances on AR_n are less than $2n\pi$. (The space AR_n is therefore bounded. It is also closed in U_n, but as it is homeomorphic to an open set in $\mathbb{R}^n$, it is not compact. On the basis of Theorem I.10.3 from Helgason (see chapter two), this also implies that U_n is not geodesically complete).

3. The map $f : AR_{n-1} \to AR_n^{\lambda=\sin(\alpha)}$ is linear: $xL_{n-1} \mapsto L_n = (I_{n+1} + \sin(\alpha)P_n)(xL_{n-1})$.
For that reason, it induces also a linear transformation f_* of the tangent space $x\Pi_{n-2}$ on AR_{n-1} in xL_{n-1} into the tangent space $x\Pi_{n-2}$ on $AR_n^{\lambda=\sin(\alpha)}$ in L_n:

$$\frac{\partial\gamma}{\partial t}_{|0} \overset{f_*}{\mapsto} \frac{\partial(f\circ\gamma)}{\partial t}_{|0} = (I_{n+1} + \sin(\alpha)P_n)(\frac{\partial\gamma}{\partial t}_{|0}). \tag{5.10}$$

The inner products on both tangent spaces have much in common. The inner product on the tangent space Π_{n-1} on AR_n in xL_{n-1} is the Schur inner product, induced by L_{n-1}. The inner product on the tangent space Π_{n-1} on AR_n in L_n is induced by $\cos(\alpha)L_{n-1}$, as we have seen above. Hence we have:

$$\langle v, w\rangle_{|L_n} = \frac{1}{\cos^2(\alpha)}\langle v, w\rangle_{|xL_{n-1}} \text{ for all polynomials } v, w \in \Pi_{n-1}. \tag{5.11}$$

From (5.10, 5.11) follows the formula of the proposition. Note further that $\tilde{\rho}(d_1, \ldots, d_{n-1}, \alpha) = f \circ \rho(d_1, \ldots, d_{n-1})$, so $\frac{\partial}{\partial d_i}_{|L_n} = f_*(\frac{\partial}{\partial d_i}_{|xL_{n-1}})$, which explains equation (5.7).

4. We already proved that $\frac{\partial}{\partial\alpha}$ is perpendicular to any submanifold $N = AR_n^{\lambda=\sin(\alpha)}$, and has length one. Hence the map $N_q \to M_q : X(q) \mapsto \nabla_X \frac{\partial}{\partial\alpha}$ is in fact Weingarten's map from N_q to N_q (see chapter two, subsection about submanifolds). As this map is symmetric, we have:

$$\langle \nabla_{\frac{\partial}{\partial d_i}} \frac{\partial}{\partial\alpha}, \frac{\partial}{\partial d_j} \rangle = \langle \nabla_{\frac{\partial}{\partial d_j}} \frac{\partial}{\partial\alpha}, \frac{\partial}{\partial d_i} \rangle,$$

and consequently:

$$\frac{\partial}{\partial\alpha}\langle \frac{\partial}{\partial d_i}, \frac{\partial}{\partial d_j} \rangle = \langle \nabla_{\frac{\partial}{\partial\alpha}} \frac{\partial}{\partial d_i}, \frac{\partial}{\partial d_j} \rangle + \langle \frac{\partial}{\partial d_i}, \nabla_{\frac{\partial}{\partial\alpha}} \frac{\partial}{\partial d_j} \rangle = 2\langle \nabla_{\frac{\partial}{\partial d_i}} \frac{\partial}{\partial\alpha}, \frac{\partial}{\partial d_j} \rangle.$$

Now, when we differentiate equation (5.7) w.r.t. α, we get the same right hand side as when we replace $\frac{\partial}{\partial d_i}$ by $\frac{1}{\cos(\alpha)} P_n(\frac{\partial}{\partial d_i})$ and multiply by 2, hence equation (5.8). The second fundamental form of the submanifold AR_{n-1} is, consequently, non-zero. Therefore the geodesics on AR_{n-1} need not be the geodesics on AR_n (see chapter two, subsection about submanifolds). □

The manifold AR_n can be parametrized by the Schurparameters $(\lambda_1, \cdots, \lambda_n) \in (-1,1)^n$, as is clear from section 2.5 in chapter two. Proposition 5.3 (and equation (5.9)) implies that on the submanifold AR_k of AR_n, the vector fields $\frac{\partial}{\partial\lambda_k}, \cdots, \frac{\partial}{\partial\lambda_n}$ are mutually perpendicular and perpendicular to the other Schur vector fields $\frac{\partial}{\partial\lambda_1}, \cdots, \frac{\partial}{\partial\lambda_{k-1}}$. In particular, the MLE of the nth Schur parameter of an AR(n) process is asymptotically uncorrelated with the MLE of the other Schur parameters. Another word for Schur parameters is partial autocorrelations. The fact that the MLE of the nth partial autocorrelation of an AR(n) process is uncorrelated with the MLEs of the first $n-1$ seems to have firstly been derived by Quenouille. A different derivation of this fact can be found in Barndorff-Nielsen and Schou (1973), also different from our derivation which was done without knowing this article. The use of the partial autocorrelations for order estimation is well known. The Wald test statistic W for testing H_0: the order of the autoregresive process $(y_t)_{0\leq t\leq T}$ is $n-1$, against H_1: the order is n, is simply $W = T\hat{\alpha}_n^2$, where $\hat{\alpha}_n$ is the MLE of α_n, the arc sine of the last Schur parameter if one uses this covariance stabilizing parametrization.

Proposition 5.3 has a number of corollaries. Firstly, we are able to calculate the projection of any polynomial $L_n \in U_n$ on U_{n-1}. Secondly, we have an example of a two-dimensional totally geodesic distribution of curvature zero. Thirdly, we are able to calculate the sectional curvature of a number of tangent planes to U_n. Fourthly, we are able to calculate all the geodesics in U_2.

Corollary 5.5 For any $L_n = xL_{n-1} + \sin(\alpha(0))E_{n-1} \in U_n$, the polynomial xL_{n-1} is the nearest point on U_{n-1} in the sense of the asymptotic Fisher metric. The distance of L_n to U_{n-1} is equal to $|\alpha(0)|$, where $\sin(\alpha(0))$ is the last Schur parameter of L_n and $\alpha(0) \in (-\frac{\pi}{2}, \frac{\pi}{2})$.

Proof:

Consider an arbitrary C^∞-curve $t \mapsto \gamma(t)$ from the polynomial L_n to the hypersurface U_{n-1}:

$$\gamma(0) = L_n \tag{5.12}$$
$$\gamma(1) \in U_{n-1} \tag{5.13}$$

Using the new parametrization, we have:

$$\gamma(t) = (d_1(t), \ldots, d_{n-1}(t), r(t), \alpha(t)),$$

and

$$\frac{\partial \gamma}{\partial t}(t) = d_1'(t)\frac{\partial}{\partial d_1} + \ldots d_{n-1}'(t)\frac{\partial}{\partial d_{n-1}} + r'(t)\frac{\partial}{\partial r} + \alpha'(t)\frac{\partial}{\partial \alpha}.$$

From proposition 5.3, we see that:

$$||\frac{\partial \gamma}{\partial t}(t)|| \geq |\alpha'(t)|.$$

If we denote the length of the curve from L_n to U_{n-1} by $|\gamma|$, then we have:

$$|\gamma| = \int_0^1 ||\frac{\partial \gamma}{\partial t}(t)||dt \geq$$

$$\geq \int_0^1 |\alpha'(t)|dt \geq |\int_0^1 \alpha'(t)dt| = |\alpha(1) - \alpha(0)| = |\alpha(0)|.$$

The right hand side is exactly the length of the curve $\alpha \mapsto xL_{n-1} + \sin(\alpha)E_{n-1}$ with starting-point L_n and end-point xL_{n-1} on U_{n-1}.
This shows that $xL_{n-1}(x)$ is the Fisher optimal AR(n-1)- approximation of L_n. □

Remark 5.6 Projecting L_n on U_{n-1} means minimizing the Riemannian distance of L_n to points in U_{n-1}. Instead of minimizing the Riemannian distance, one could minimize the Kullback-Leibler divergence of L_n to points in U_{n-1}. As Amari has shown, also the Kullback-Leibler divergence is based on geometry, derived from the asymptotic Fisher matrix as Riemannian metric. This geometry uses the 1- and -1-connections, instead of the 0- or Levi-Civita-connection that we use. Then one finds $\cos(\alpha)xL_{n-1}(x)$, instead of $xL_{n-1}(x)$ as nearest point of L_n on U_{n-1}.
Another principle of projecting L_n on U_{n-1} is maximization of the entropy over all AR(n-1) processes with same autocovariances $\eta_0, \cdots, \eta_{n-1}$ as the AR(n)-process L_n. Again, one finds $\cos(\alpha)xL_{n-1}(x)$ as the point in U_{n-1} which is nearest to L_n in this respect. See Amari (1987) and Landau (1987).
The strong advantage of the Kullback-Leibler divergence in comparison with the Riemannian distance is that one can also project L_n easily on U_{n-k} for all $1 \leq k \leq n$. It is just a repeated application of the procedure with the last, second last etc. Schur parameter. This is not clear for the Riemannian distance; related with this is the fact that the Riemannian distance of U_n, restricted to U_{n-1}, is not the same as the Riemannian

distance of U_{n-1}. This is because U_{n-1} is not totally geodesic in U_n. Hence, a geodesic (or shortest curve) on U_{n-1} that connects two points in U_{n-1}, need not be a geodesic (or shortest curve) in U_n, and the U_n-distance between the two points will be smaller than the U_{n-1}-distance.
For our purpose i.e. coefficient modelling for time-varying autoregressive processes, one advantage of the Levi-Civita-connection, however, is that AR_n is totally geodesic in U_n. This is not true for other connections. As we shall show in chapter six, this makes it possible—to some extent—to separate the identification of the moving noise-levels from the identification of the (other) moving coefficients. □

The next corollary 5.7 is an illustration for proposition 3.9.

Corollary 5.7 The distribution generated by $\frac{\partial}{\partial r}$ and $\frac{\partial}{\partial \alpha}$ has a constant dimension two, is totally geodesic, and has a curvature zero.

Proof: Immediate consequence of the propositions of the last two subsections. □

The next corollary 5.8 gives a type of parametrization of U_n, converting the asymptotic Fisher matrix to a simpler form. At the same time the sectional curvature of tangent planes, containing the vector $\frac{\partial}{\partial \alpha}$, is calculated. It turns out that this sectional curvature is negative, except for the sectional curvature of the plane spanned by $\frac{\partial}{\partial r}$ and $\frac{\partial}{\partial \alpha}$, which is zero as we just saw. This has consequences for stochastic processes on U_n of independent directional variation, according to the asymptotic Fisher metric.

Corollary 5.8

1. Let $(a_1, \ldots, a_{n-1})$ be the natural parameters of AR_{n-1}. In that case, we get another parametrization $\rho(d_1, \ldots, d_{n-1})$ of AR_{n-1} by defining:

$$d_i = \frac{a_i + a_{n-i}}{\sqrt{2},} \text{ if } 1 \leq i \leq \tfrac{n}{2} \text{ and}$$

$$d_i = \frac{a_i - a_{n-i}}{\sqrt{2},} \text{ if } \tfrac{n}{2} < i \leq n-1.$$

With these parameters the metric on AR_{n-1} has the form:

$$H_\rho^{AR_{n-1}} = \begin{pmatrix} H_1 & 0 \\ 0 & H_2 \end{pmatrix} \tag{5.14}$$

Then, w.r.t. the parameters $d_1, \ldots, d_{n-1}, \alpha, r$ (where $r = \log \sigma$, as in the preceding theorem) the metric on U_n has the form:

$$H_{\tilde{\rho}}^n = \begin{pmatrix} \frac{1+\sin(\alpha)}{1-\sin(\alpha)} H_1 & 0 & 0 & 0 \\ 0 & \frac{1-\sin(\alpha)}{1+\sin(\alpha)} H_2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix} \tag{5.15}$$

Furthermore, we have:

$$\nabla_{\frac{\partial}{\partial d_i}} \frac{\partial}{\partial \alpha} = \frac{1}{\cos(\alpha)} \frac{\partial}{\partial d_i}, \text{ if } i \leq \tfrac{n}{2},$$

$$\nabla_{\frac{\partial}{\partial d_i}} \frac{\partial}{\partial \alpha} = \frac{-1}{\cos(\alpha)} \frac{\partial}{\partial d_i}, \text{ if } i > \tfrac{n}{2}, \text{ and}$$

$$\nabla_X \frac{\partial}{\partial r} = 0 \text{ for all vector fields } X.$$

2. The sectional curvature of the plane $W = [\frac{\partial}{\partial d_i}, \frac{\partial}{\partial \alpha}]$ in the tangent space in L_n to U_n is

$$K(W) = \frac{-1}{1 - \sin(\alpha)} \quad (i \leq \frac{n}{2}), \tag{5.16}$$

and

$$K(W) = \frac{-1}{1 + \sin(\alpha)} \quad (i > \frac{n}{2}). \tag{5.17}$$

Consequently, if $n > 1$, the sectional curvature of U_n is not constant.

Proof:

$$\frac{\partial}{\partial d_i} = \frac{1}{\sqrt{2}}(\frac{\partial}{\partial a_i} + \frac{\partial}{\partial a_{n-i}}) \in x\Pi_{n-2} \text{ if } \leq i \leq \tfrac{n}{2}, \text{ and}$$

$$\frac{\partial}{\partial d_i} = \frac{1}{\sqrt{2}}(\frac{\partial}{\partial a_i} - \frac{\partial}{\partial a_{n-i}}) \in x\Pi_{n-2} \text{ if } n > i > \tfrac{n}{2},$$

hence, $\frac{\partial}{\partial d_i}$ is the eigenvector of P_n, corresponding to the eigenvalue 1, if $1 \leq i \leq \frac{n}{2}$ and $\frac{\partial}{\partial d_i}$ is the eigenvector of P_n, corresponding to eigenvalue -1 if $n > i > \frac{n}{2}$. Note that the Schur and Fisher metric coincide on Π_{n-1}, and that under the Schur inner product P_n is an isometric transformation of $x\Pi_{n-2}$. This explains equation (5.14). Application of formula (5.7) and proposition 5.3 item 1 yields equation (5.15). Application of the sectional curvature formula from the preceding section yields equations (5.16, 5.17).□

Remark 5.9 From corollary 5.8, it is clear that U_2 has a non-positive curvature everywhere. It is also easy to prove that in general *any* two-dimensional tangent plane to U_n, containing the vector $\frac{\partial}{\partial \alpha}$ has non-positive curvature. But it is not true that U_n has non-positive curvature for *any* $n > 1$. Take, for instance, the plane V, generated by $\frac{\partial}{\partial a_1}, \frac{\partial}{\partial a_2}$ in the point $a_1 = 0, a_2 = \frac{1}{2}, a_3 = 0, \sigma = 1$ in U_3. V is also a tangent plane to the hypersurface in U_3 of all the polynomials with last Schur parameter equal to zero. This hypersurface is just U_2. In U_2 the sectional curvature of V is $\tilde{K}(V) = -\frac{2}{3}$, according to the above corollary. But in U_3 we have to use the formula of Gauss (2.18, see chapter two). Weingarten's map of U_2 in U_3 is just $P_3 : x\Pi_1 \to x\Pi_1$, as proposition 5.3 shows. We have:

$$-1 = \det(P_3) = \langle P_3 v, v\rangle \langle P_3 w, w\rangle - \langle P_3 v, w\rangle^2$$

for any orthonormal base v, w of $x\Pi_1$. Applying Gauss' formula (2.18, see chapter two), we obtain:

$$K(V) = \tilde{K}(V) - (\langle P_3 v, v\rangle \langle P_3 w, w\rangle - \langle P_3 v, w\rangle^2) = \frac{1}{3}.\square$$

We now come to an important fact. By the observation about the curvature we can prove that there is *no covariance stabilizing parametrization* of the whole model set of AR(n) processes if $n > 1$.

Proposition 5.10 It is not possible to find a parametrization $\chi : V \subset \mathbf{R}^{n+1} \to U_n; v \mapsto \chi(v)$ of U_n, such that the asymptotic Fisher matrix $H_\chi(v)$ with respect to this parametrization does not depend on v if $n > 1$.

Proof: If $H_\chi(v)$ does not depend on v, then $\nabla_{\frac{\partial}{\partial v_i}} \frac{\partial}{\partial v_j} = 0$ for all $1 \leq i, j \leq n+1$, thus the sectional curvature would be zero for any plane in any tangent space. This contradicts the preceding corollary 5.10. □

5.2 The asymptotic Fisher metric on U_2

In this section we study the coefficient space for stationary AR(2) processes in more detail.
Corollary 5.8 shows a simpler form of the asymptotic Fisher information matrix with respect to the Schur parameter parametrization. In case of U_2, this simpler form makes it easy to calculate all the geodesics.

Corollary 5.11 If β, α are the arcsine of the first and last Schur parameters of polynomials in U_2 and $r = \log(\sigma)$, then w.r.t. the parameters β, α, r these polynomials have the form:

$$L_n(x) = e^{-r}(x^2 + \sin\beta(1+\sin\alpha)x + \sin\alpha) \quad (L_n \in U_2) \tag{5.18}$$

and the Fisher metric on U_2 is:

$$H^2_{\beta,\alpha,r} = \begin{pmatrix} \frac{1+\sin\alpha}{1-\sin\alpha} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} \tag{5.19}$$

The equations for geodesics $\gamma(t) = (\beta(t), \alpha(t), r(t))$ parametrized, according to the arc-length are:

$$\frac{d^2\beta}{dt^2} + \frac{2}{\cos\alpha}\frac{d\beta}{dt}\frac{d\alpha}{dt} = 0 \tag{5.20}$$

$$\frac{d^2\alpha}{dt^2} - \frac{1}{\cos\alpha}\frac{1+\sin\alpha}{1-\sin\alpha}(\frac{d\beta}{dt})^2 = 0 \tag{5.21}$$

$$\frac{d^2r}{dt^2} = 0 \tag{5.22}$$

$$\frac{1+\sin\alpha}{1-\sin\alpha}(\frac{d\beta}{dt})^2 + (\frac{d\alpha}{dt})^2 + 2(\frac{dr}{dt})^2 = 1 \tag{5.23}$$

A formula for all the geodesics through a point $(\alpha(0), \beta(0), r(0))$ in U_2 is:

$$\sin\alpha(t) = \sin^2\gamma + \cos^2\gamma \sin(qt+\delta); \tag{5.24}$$

$$\beta(t) = \beta(0) - 2\arctan\left(\frac{\cos\delta}{\sin\gamma(1-\sin\delta)}\right) - \sin\gamma \times qt + 2\arctan\left(\frac{\cos(qt+\delta)}{\sin\gamma(1-\sin(qt+\delta))}\right) \text{ if } \gamma \neq 0; \tag{5.25}$$

$$\beta(t) = \beta(0) \text{ if } \gamma = 0; \tag{5.26}$$

$$r(t) = r(0) \pm \sqrt{\frac{1-q^2\cos^2\gamma}{2}} \times t \tag{5.27}$$

where:

- $-\frac{\pi}{4} - \frac{\alpha(0)}{2} \leq \gamma \leq \frac{\pi}{4} + \frac{\alpha(0)}{2}$;
- $\delta \in [-\frac{\pi}{2}, \frac{\pi}{2})$ is a solution of $\sin(\alpha(0)) = \sin^2\gamma + \cos^2\gamma \sin\delta$;
- $q^2\cos^2\gamma \leq 1$;
- $\beta(0) \in (-\frac{\pi}{2}, \frac{\pi}{2})$;
- $r(0) \in \mathbb{R}$, and
- t is an element of the largest interval around 0, such that $|\sin\alpha(t)| \neq 1$ and $\beta(t) \in (-\frac{\pi}{2}, \frac{\pi}{2})$ (if $q \neq 0$ this interval is bounded).

If $q = 0$, then we have the geodesics of the noise pencil, treated in this section, and if $q^2 = 1$, $\gamma = 0$, we have the geodesics of the Schur pencil.
All geodesics have a finite maximal interval, except the geodesics of the noise pencil, i.e. the geodesics, described by the log noise level.
The lengths of the maximal intervals of the geodesics on AR_2 are uniformly bounded.

Proof:
The form of the matrix $H^2_{\beta,\alpha,r}$ is a direct consequence of proposition 5.3 item 2. We know already that $\nabla_{\frac{\partial}{\partial\beta}}\frac{\partial}{\partial\alpha} = \frac{1}{\cos\alpha}\frac{\partial}{\partial\beta}$, and it is also easy to calculate that $\nabla_{\frac{\partial}{\partial\beta}}\frac{\partial}{\partial\beta} = -\frac{1}{\cos\alpha}\frac{1+\sin\alpha}{1-\sin\alpha}\frac{\partial}{\partial\alpha}$. From this the remainder can be derived (the equation (5.23) expresses that we want to parametrize, according to arc length). □

This corollary 5.11 has a consequence for stochastic processes on totally geodesic submanifolds M of U_2 of independent directional variation, according to the asymptotic Fisher metric of the General Model type.

Corollary 5.12 If the noise level σ on the open submanifold M of U_2 is bounded away from zero and from infinity, or if the submanifold M is equal to $\frac{1}{\sigma}AR_2$, then the running

time is a.s. finite.

Proof:
If the submanifold M is open or $M = \frac{1}{\sigma}AR_2$, then M is totally geodesic. Because of the bound on the noise level, the lenghts of the maximal domains of the geodesics on M are uniformly bounded. Application of lemma 3.16 ends the proof. □

We conclude this section by showing that U_2 is a normal neighbourhood of any of its points. The proof, by elementary means, is based on showing that any two points in U_2 can be connected by one and only one geodesic.

Corollary 5.13 U_2 is a normal neighbourhood of any of its points. Any two points in U_2 can be connected by one and only one geodesic.

The proof is structured as follows. First, we prove that the projections on $\frac{1}{\sigma}AR_2$ of the geodesics are also geodesics, and except for the geodesics of the Schur pencil ($\gamma = 0$), they satisfy the following equations:

$$(\frac{da_2}{d\beta})^2 = \cot^2(\gamma)\frac{(1-a_2)^3}{(1+a_2)} - (1-a_2)^2 \tag{5.28}$$

$$\beta = \beta_0 \pm (2\arctan(\frac{1}{\sin\gamma}\sqrt{\frac{\cos(2\gamma)-a_2}{1+a_2}}) - 2\sin\gamma \times \arctan(\sqrt{\frac{\cos(2\gamma)-a_2}{1+a_2}})) \tag{5.29}$$

$$\gamma \in (0, \frac{\pi}{2}) \tag{5.30}$$

From this we shall conclude the following:

1. geodesics on $\frac{1}{\sigma}AR_2$ of the form (5.29) with the same β_0, but with different $\gamma > 0$, do not intersect;
2. two geodesics on U_2 have at most one point of intersection;
3. the union of the geodesics on $\frac{1}{\sigma}AR_2$ of the form (5.29) with the same β_0 is equal to $\frac{1}{\sigma}AR_2$;
4. any two points in U_2 can be connected by a geodesic.

Proof:

The Fisher metric on $\frac{1}{\sigma}AR_2$ w.r.t. the parameters β, α is:

$$H^2_{\beta,\alpha} = \begin{pmatrix} \frac{1+\sin\alpha}{1-\sin\alpha} & 0 \\ 0 & 1 \end{pmatrix}, \tag{5.31}$$

thus, it is possible to extend the Riemannian manifold $\frac{1}{\sigma}AR_2$ to $M = (-\infty, \infty) \times (-1, 1)$ with parameters β and $a_2 = -\sin(\alpha)$. With respect to these parameters the geodesics of M of the Schur pencil ($\gamma = 0$) are the vertical lines. Furthermore, have geodesics of

M of the form (5.29) the following properties: they are the graphs of concave functions with top $(\beta_0, \cos(2\gamma))$, domain $(\beta_0 - (1-\sin\gamma)\pi, \beta_0 + (1-\sin\gamma)\pi)$, and axis of symmetry at $\beta = \beta_0$. Note that a higher top corresponds to a larger domain.
Now, assume that two geodesics of the form (5.29) with the same β_0, but different γ intersect. In that case, because of symmetry, they would intersect as well in the increasing part of the graph as in the decreasing part. However, from the remark made above (on top and domain), it is clear that they have to intersect at least twice in the increasing part. Since on the increasing part one can consider β as a function of a_2 for both curves, e.g. $\beta(a_2)$ and $\tilde{\beta}(a_2)$ respectively, the difference $\frac{d\beta}{da_2} - \frac{d\tilde{\beta}}{da_2}$ must be zero for some a_2, which contradicts (5.28).
Almost the same reasoning can be applied to prove that two geodesics on M have one point of intersection at most. Vertical geodesics intersect other geodesics in at most one point, because the latter are graphs. Consequently, it suffices to look to two geodesics of type (5.29). Suppose that they have two points of intersection. It is not possible that these two points are in the increasing part of both graphs, or that both points are in the increasing part of one and in the decreasing part of the other, just because of the same argument as above. Suppose that one point of intersection is in the increasing part of both geodesics and the other in the decreasing part of one of them. Then, the domain of one of them must be smaller than and contained in the other domain. However, equation (5.28) shows that its top is higher than that of the other, which is impossible. A point of intersection in the increasing part of one and the decreasing part of the other and another point of intersection vice versa is not possible, because the decreasing part is on the right hand for all geodesics of form (5.29). This proves the second statement as far as M and therefore $\frac{1}{\sigma}AR_2$ are concerned. To prove the second statement for U_2, we suppose that two geodesics in U_2 intersect in two points. We project the two geodesics on AR_2. The proof follows now from the observation that the projection of a geodesic on AR_2 is injective and again a geodesic unless it is given by $r(t) = r(0) \pm t$, with α and β constant, in which case the projection only consists of one point.
For the third statement, we show that if

$$V = \{(\gamma, a_2) | 0 < \gamma < \frac{\pi}{2}, -1 < a_2 \leq \cos(2\gamma)\} \text{and}$$

$$f : V \mapsto [\beta_0, \beta_0 + \pi) \times (-1, 1) \text{ is defined by:}$$

$$f(\gamma, a_2) = (\beta_0 + (2\arctan(\frac{1}{\sin\gamma}\sqrt{\frac{\cos(2\gamma) - a_2}{1 + a_2}}) -$$

$$- 2\sin\gamma \times \arctan(\sqrt{\frac{\cos(2\gamma) - a_2}{1 + a_2}})), a_2) \tag{5.32}$$

then, f is surjective. This follows directly from the observation that, for any $a_2 \in (-1, 1)$

$$f(\arccos(a_2)/2, a_2) = 0; \lim_{\gamma \to 0} f(\gamma, a_2) = \pi,$$

and the map $\gamma \mapsto f(\gamma, a_2)$ is continuous.

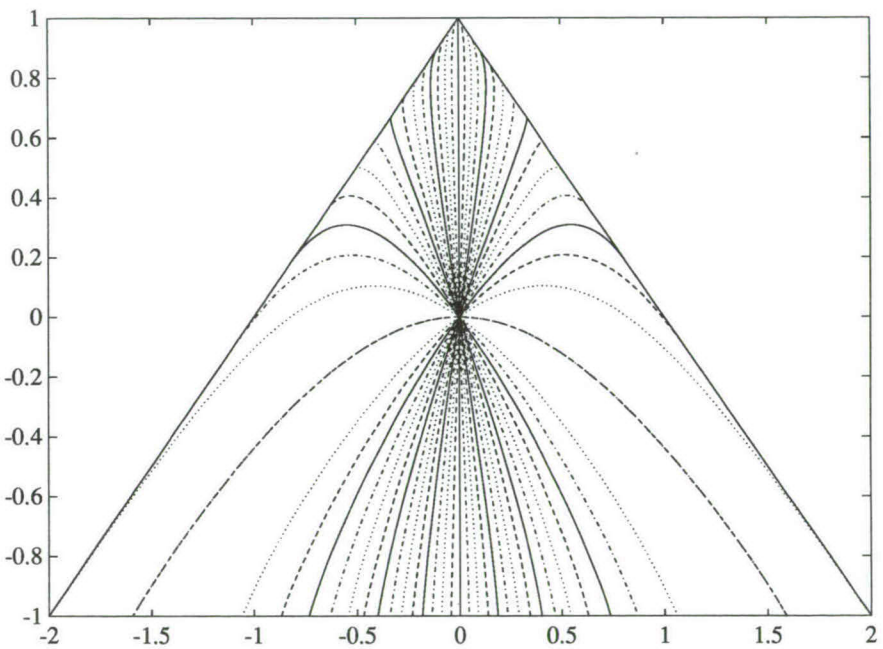


Figure 5.1: pencil of geodesics in the asymptotic Fisher metric through the point $(0,0)$ in AR_2

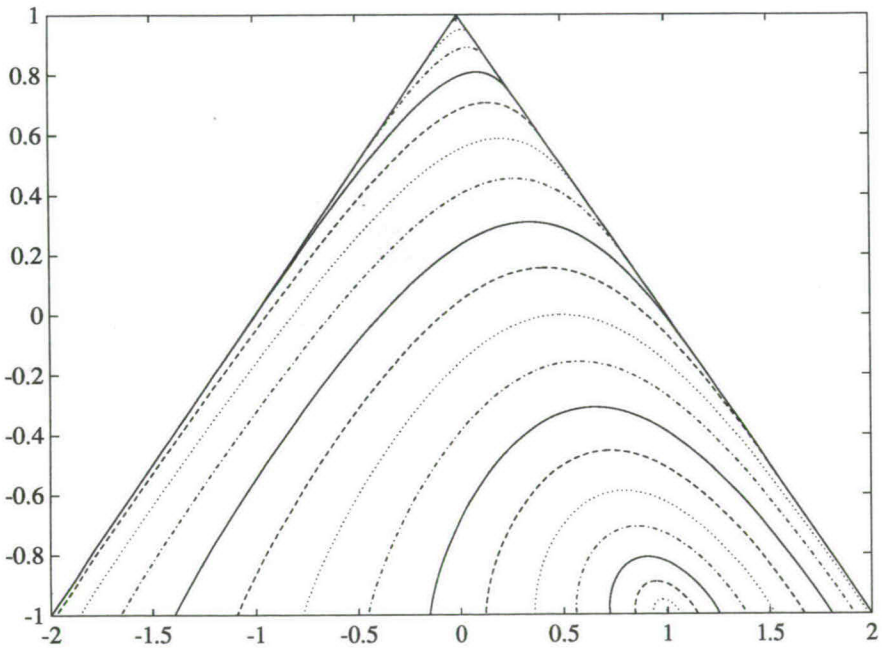


Figure 5.2: other pencil of geodesics in the asymptotic Fisher metric on AR_2

As $\frac{1}{\sigma}AR_2$ is the part of M, corresponding to $(-\frac{\pi}{2},\frac{\pi}{2})\times(-1,1)$, we have shown that the part of the geodesics at the right hand side of the vertical line $\beta=\beta_0$ covers the right hand side of $\frac{1}{\sigma}AR_2$. In the same way, one proves that the left hand side of the geodesics covers the left hand side of $\frac{1}{\sigma}AR_2$.
We prove the fourth statement first on $\frac{1}{\sigma}AR_2$. Without loss of generality, we restrict to the case that one of the two points to be connected is $\beta=0; a_2=0$. Introduce the expressions:

$$b(\gamma) = -2\sin\gamma \times \arctan(\sqrt{\cos(2\gamma)}) + 2\arctan(\frac{1}{\sin\gamma}\sqrt{\cos(2\gamma)}) \quad (5.33)$$

$$k(\gamma,a_2) = 2\arctan(\frac{1}{\sin\gamma}\sqrt{\frac{\cos(2\gamma)-a_2}{1+a_2}}) - 2\sin\gamma\times\arctan(\sqrt{\frac{\cos(2\gamma)-a_2}{1+a_2}}) \quad (5.34)$$

and the set:

$$V = \{(\gamma,a_2)|0<\gamma\leq\frac{\pi}{4}; -1<a_2\leq\cos(2\gamma)\} \quad (5.35)$$

Further, we define the functions:

$$g_1 : V\to M; g_1(\gamma,a2) = (b(\gamma)-k(\gamma,a_2),a_2) \quad (5.36)$$

$$g_2 : V\to M; g_1(\gamma,a2) = (b(\gamma)+k(\gamma,a_2),a_2) \quad (5.37)$$

$$g_3 : V\to M; g_1(\gamma,a2) = (-b(\gamma)+k(\gamma,a_2),a_2) \quad (5.38)$$

$$g_4 : V\to M; g_1(\gamma,a2) = (-b(\gamma)-k(\gamma,a_2),a_2) \quad (5.39)$$

For any $\gamma\in(0,\frac{\pi}{4})$, the curve $a_2\mapsto g_i(\gamma,a_2)$ is the increasing or decreasing part of a (not on arc-length parametrized) geodesic through $(0,0)$ with top $(b(\gamma),\cos(2\gamma))$ if $i\leq 2$ and with top $(-b(\gamma),\cos(2\gamma))$ if $i\geq 3$. Note that

$$\lim_{\gamma\to 0} b(\gamma) = \pi \quad (5.40)$$

$$\lim_{\gamma\to 0} k(\gamma,a_2) = \pi\ \forall a_2\in(-1,1) \quad (5.41)$$

With this, and a continuity argument, it is easy to show that the union of the vertical geodesic through $(0,0)$ and the images of the functions g_i is equal to $(-2\pi,2\pi)\times(-1,1)$, which proves the fourth statement as far as $\frac{1}{\sigma}AR_2$ is concerned. Now, if two points in U_2 are given, we first connect their projections on AR_2 by a geodesic, and thereupon, we only have to readjust the values for $r(0)$ and q in the equations (5.24) to find the geodesic through the two points in U_2.

5.3 Other metrics on U_n

In this section we look for other metrics on U_n, satisfying some desirable properties. The asymptotic Fisher metric on U_n is such a metric. It has the property that $\nabla_X\frac{\partial}{\partial r}=0$ for all vector fields X on U_n. This property makes it possible to separate identification of noise-levels from identification of other coefficients, to some extent, as will be shown in chapter six. We shall say that manifolds and metrics with this property satisfy the *separation criterion.*

Definition. A manifold M satisfies the separation criterion if

- either M is open in U_n or M is open in $\frac{1}{\sigma}AR_n$;
  the metric on M is induced by a metric on an open part $\tilde{M}$ of U_n.

- the log-noise vector field $\frac{\partial}{\partial r}$ and the coefficient vector fields $\frac{\partial}{\partial a_i}$ $(1 \leq i \leq n)$ satisfy

$$\nabla_X \frac{\partial}{\partial r} = 0 \qquad \forall X \in \mathcal{D}\tilde{M} \tag{5.42}$$

  and

$$\langle \frac{\partial}{\partial r}, \frac{\partial}{\partial a_i} \rangle = 0. \tag{5.43}$$

- Furthermore, the log-noise-levels have to be bounded: we assume that there exists a number $\mathcal{M} > 0$, such that

$$|r(q)| < \mathcal{M} \quad \forall q \in M. \qquad \square \tag{5.44}$$

The assumption on the log-noise vector field (5.42) implies that the inner products

$$\langle \frac{\partial}{\partial a_i}, \frac{\partial}{\partial a_j} \rangle; \qquad \langle \frac{\partial}{\partial r}, \frac{\partial}{\partial a_j} \rangle; \qquad \langle \frac{\partial}{\partial r}, \frac{\partial}{\partial r} \rangle$$

are constant along r-curves. Hence, the metric has a matrix on the tangent spaces with respect to the natural coordinate base of the form:

$$H_{nat}(a_1, \cdots, a_n, r) = \begin{pmatrix} H(a_1, \cdots, a_n) & 0_{n\times 1} \\ 0_{1\times n} & c \end{pmatrix};$$

a matrix that does not depend on r. The matrix H defines mutually isometrically isomorphic Riemannian structures on the manifolds $\frac{1}{\sigma}AR_n \cap M$ for all noise levels σ^2. Furthermore, the r-curves are geodesics. The projection of any geodesic in M on the manifold $M \cap \frac{1}{\sigma}AR_n$ is again a geodesic or a single point.

If the separation criterion is satisfied, it is possible to model the process of the instantaneous associated polynomials $L_{n,t}(x) = e^{-r_t}(x^n - a_{1,t}x^{n-1} - \cdots - a_{n,t})$, according to the General Model, by modelling *two processes independently from each other*:

- the process of the *projected* instantaneous polynomials $x^n - a_{1,t}x^{n-1} + \cdots - a_{n,t}$, according to the General Model on AR_n;

- and the process of the log-noise-levels, according to the General Model on $\mathbb{R}$.

If the process of the projected polynomials on AR_n is modelled according to a zero-curvature model, the complete model of the combined processes is also a zero-curvature model.
An important example of such a zero-curvature model arises if the process of the projected polynomials is modelled according to a Geodesic Model on AR_n. Then, we call the complete model a *Geodesic plus Noise* Model. In terms of "direction spaces" (see

chapter three), we can define this two-dimensional model formally as follows.

Definition. A Geodesic plus Noise Model is a General Model on a manifold that satisfies the separation criterion, such that the direction spaces $\mathcal{L}_t$ all satisfy

$$\mathcal{L}_t = \mathcal{L} = \boldsymbol{span}\{F, \frac{\partial}{\partial r}\},$$

where F is a geodesic tangent field perpendicular to the log-noise vector field $\frac{\partial}{\partial r}$. □

5.3.1 Complete metrics on U_n

We give an example of a complete metric on U_n. It is not difficult to construct complete metrics. In fact, there is much freedom in choosing a particular complete metric. Therefore we formulate some additional features we would like to be satisfied by the complete metric in view of the application to the General Models.

1. The metric should satisfy the separation criterion. Consequently, the submanifolds $\frac{1}{\sigma}AR_n$ will be totally geodesic, and have mutually isometric Riemannian structures.

2. The metric should have zero-curvature. This will enable us to apply identification methods developed for zero-curvature models, presented in chapter six, to models of various dimensions.

3. Using the criterion of the Kullback-Leibler divergence, the projection of the polynomial $L_n(x) \in \frac{1}{\sigma}AR_n$ in U_n on U_{n-k} is $\sqrt{1-\lambda_n^2}\cdots\sqrt{1-\lambda_{n-k+1}^2}x^k L_{n-k}(x)$ (notations as in chapter two, section 2.5.).
   Using the criterion of the Riemannian distance given by the asymptotic Fisher metric, the projection of $L_n(x)$ on U_{n-1} will be $xL_{n-1}(x)$ on the same submanifold $\frac{1}{\sigma}AR_n$.
   Also, if we use the metric that we are constructing as a criterion, the projection of $L_n(x) \in \frac{1}{\sigma}AR_n$ on U_{n-k}, will be on the same manifold $\frac{1}{\sigma}AR_n$, because $\frac{1}{\sigma}AR_n$ will be totally geodesic. In order that the metric retains some relation with the statistics of time-invariant autoregressive processes, we propose that the projection of $L_n(x)$ on U_{n-k} should be $x^k L_{n-k}(x)$.

Let $\tilde{\Lambda} : (-1,1)^n \to AR_n$; $(\lambda_1, \cdots, \lambda_n) \mapsto L_n(x)$ be the parametrization of the manifold of monic stable polynomials by the Schur parameters, as introduced in chapter two, section 2.5 . Then, the requirements 1., 2. and 3. will be realized if the matrix of the metric on the parametrization

$$\Lambda : (-1,1)^n \times \mathbb{R} \to U_n, \quad \Lambda(\lambda_1, \cdots, \lambda_n, \rho) = e^{-\sqrt{2}\rho} L_n(x)$$

has the form:

$$H_\Lambda(\lambda_1,\cdots,\lambda_n,r)=\begin{pmatrix} f_1(\lambda_1) & 0 & \cdots & 0 & 0 \\ 0 & f_2(\lambda_2) & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & f_n(\lambda_n) & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}. \tag{5.45}$$

We introduce the parameters β_i defined by:

$$\beta_i:(-1,1)\to V_i;\quad \beta_i(\lambda_i)=\int_0^{\lambda_i} f_i^{\frac{1}{2}}(x)dx. \tag{5.46}$$

With respect to the parameters $(\beta_1,\cdots,\beta_n,\rho)$, the metric has a matrix equal to the identity. This shows that the curvature is zero. It is also immediately clear that the metric wil be complete if and only if $V_i=\mathbb{R}$. For example, if the function $f_i:(-1,1)\to\mathbb{R}^+$ has the form $f_i(x)=f(x)=\frac{1}{(1-x^2)^h}$, then this will be the case if and only if $h\geq 2$. For $h=2$, this yields

$$\lambda_i=\tanh(\beta_i). \tag{5.47}$$

We have seen for the set of stationary AR(1) models that this parametrization yields an ML estimate of β_1, such that the asymptotic skewness of its distribution is zero.

5.3.2 The Euclidean metric on U_n

Let AR_n be the set
$AR_n=\{(a_1,\cdots,a_n)|x^n-a_1x^{n-1}-\cdots-a_n$ is stable $\}$. The "natural" parametrization $\chi:AR_n\times\mathbb{R}\to U_n$ will be defined by $\chi(a_1,\cdots,a_n,\rho)=e^{-\sqrt{2}\rho}(x^n-a_1x^{n-1}-\cdots-a_n)$. The Euclidean metric will be defined as the Riemannian metric, such that its matrix with respect to the parametrization χ is the identity. Then, also this metric satisfies the separation criterion.

Chapter 6

Identification

This chapter is dedicated to identification techniques for time-varying autoregressive processes. Given a sequence of data coming from such a process, we want to describe the underlying process of coefficients and noise levels. We assume that we already have selected a class of General Models that is suitable for this purpose. Our task will be twofold.

- *Identification of hyperparameters*: We have to select a specific General Model from the class that is optimal, in some sense, for helping us to describe the coefficient process. This selection will be done by computing optimal values for the hyper-structural parameters that determine the individual elements of the class.
- *Identification of coefficients and noise-levels*: A General Model helps us in describing the coefficient process by giving a prior distribution to the coefficients and noise levels. Once a General Model has been selected from the class, it will be our task to compute the posterior distribution, given the data.

The main emphasis will be on classes of zero-curvature models. Zero-curvature models were defined in chapter three. There are charts which cover the manifold, such that the transition equation of the zero-curvature model is linear with respect to these charts. Important examples of zero-curvature models are geodesic models and geodesic plus noise models. Geodesic models exist for all manifolds and metrics. At the end of the chapter, we shall also discuss the simplest type of non-zero-curvature models.

First, we shall discuss the identification of coefficients and noise levels if all hyperstructural parameters are given. The determination of the posterior distribution of coefficients and noise levels will be studied. For the usual random walk model for the coefficient process the posterior distribution of the coefficients is Gaussian, hence, in that case it suffices to compute the conditional expectation and the conditional variance of the coefficients, given the data. This can be done by application of the Kalman Filter followed by Kalman Smoothing. The conditional expectation coincides with the maximum location of the posterior density of the coefficients. However, if we do not use the Euclidean metric, our models are non-linear in general and the posterior distribution will not be Gaussian anymore. But, if the posterior distribution does not divert too much from normality, the maximum of the posterior density is still a paramount distinguishing mark of

this distribution. Hence, a first step in identifying the coefficients will be maximization of their posterior density, or, as it is called, posterior mode estimation.
Maximization of the posterior density of the coefficients is the same as maximizing the joint density of coefficients and observations with respect to the coefficients. This gives a criterion with a nice interpretation. Minus the logarithm of the joint density is the sum of squared fitting errors plus a sum of penalties for changing a coefficient into the subsequent one. This spline interpretation of smoothing goes back to Whittaker (1923), and has again been pointed out by Kitagawa and Gersch (1985) and others. In this criterion the chosen Riemannian metric on the coefficient space acts as a set of weights in the sum of the transition penalties.
Another important feature of the posterior distribution is the concentration of the posterior density around its maximum or "mode". This can be measured by the second derivative $\bar{U}$ of minus the logarithm of the density in the maximum point. In the Euclidean case the inverse of this second derivative is the conditional variance of the coefficients, given all data. Hence, if the posterior distribution is not far from normality, one can conceive of the diagonal elements of the inverse of $\bar{U}$ as error variances of the coefficient estimates.
We shall restrict ourselves to the discussion of algortihms that approximate these two basic features of the posterior distribution: its mode and the matrix $\bar{U}$.

The methods used to maximize the posterior or joint density with respect to the coefficients (and noise levels) will be variations of the Newton search routine. For zero-curvature models it will be possible to apply a Gauss-Newton iterative procedure. The procedure we shall propose is based on an article of Fahrmeir and Kaufmann (1991). Generalizing this article slightly, in such a way that it is applicable to the situation we are studying, we shall show that for zero-curvature models Extended Kalman Filtering and Smoothing can be regarded as a simplification of this Gauss-Newton procedure. Moreover, it will be shown that for the manifolds we are studying, this variant of the Newton search routine implies a separation of every iteration step into two parts: one for updating the coefficients, the other for updating the noise levels. It will also be shown that it is not necessary for identification to have *analytic* knowledge of the parametrizations, on which the transition equation of the zero-curvature model is completely linear. An approximation method for these parametrizations will be given.

Secondly, we shall discuss the identification of hyperparameters. This will be done for the geodesic and (heuristically) for the geodesic plus noise model. The criterion for optimalisation of the values for the hyperparameters will be aiming at the maximum of the likelihood. For obtaining this maximum, Gauss-Newton and Expectation-Maximization algorithms are used.
Finally, we discuss the identification of Simple Models. In chapter three, we made a list of complexity types, and the Simple Model was then seen as the least complex model. Yet, identification is not that simple if the model is a non-zero-curvature one. The Gauss-Newton search routine for the posterior mode is now much farther away from Extended Kalman Filtering and Smoothing than in the zero-curvature models.
The chapter will be concluded by showing results of experiments where, on simulated

data series, we applied the geodesic model and the most simple model, using the Euclidean and also the asymptotic Fisher metric.

We make the following assumptions on the models for which we are going to discuss identification techniques.

1. the manifold M satisfies the separation criterion (see chapter five, section 5.3);

2. the process of the projections $x^n - a_{1,t}x^{n-1} - \cdots - a_{n,t}$ of the instantaneous polynomials on AR_n and the process of the log-noise-levels $r_t \in \mathbf{R}$ are modelled as independent processes;

3. the direction space $\mathcal{L}$ and the symmetric (or variance) tensor field V are constant in time; (see for these concepts chapter three, section 3.5)

4. the probability distribution of the first instantaneous associated polynomial q_0 is given by the transition equation from a fixed point $q_{-1} \in U_n$, according to a Special Model with direction space $\mathcal{L}$;

5. the stationary process θ, used to generate the instantaneous associated polynomials, is an autoregressive process of order zero (white noise; in Special Model), or of order one (just as we used in the sequence of General Models that admitted the time-varying spectral density).

6.1 Posterior mode estimation

We shall study identification of the coefficients and noise levels by posterior mode estimation, using modified Newton search routines. We shall assume that all hyperparameters are given.

In 6.1.1, we show that posterior mode estimation boils down to minimizing a criterion which is just minus the logarithm of the joint probability density of coefficients, noise levels and observations. We give (approximate) expressions for this criterion for all general models satisfying the assumptions 1,2,3,4,5 mentioned above.

Then, in 6.1.2 we elaborate the Newton iteration step for *zero-curvature models (e.g. geodesic models)*by calculating both the derivative and second derivative of the criterion.

In 6.1.3, we modify the gain matrix of the Newton step. By taking conditional expectations, the gain matrix is modified in such a way that the corresponding iteration step can be separated in a half step to update the coefficients, and another half step to update the noise levels. The step to update the coefficients is then just the Gauss-Newton iteration step of a non-linear least squares problem.

We took the idea to do posterior mode estimation for zero-curvature models by means of a (Gauss)-Newton search routine from an article of Fahrmeir and Kaufmann (1991).

In subsections 6.1.4 and 6.1.5, where we consider the updating of the coefficients, we

mainly elaborate that article, and adapt it for our purposes. In this article, posterior mode estimation in the theory of Dynamic Exponential Family Regression is treated. This theory deals with models of the following form:

- the observations y_t have a conditional distribution given the former observations $y_0, \cdots, y_{t-1}$ belonging to an exponential family parametrized by $\beta \in \mathbb{R}^p$, e.g. a conditional Gaussian distribution with mean depending on $y_0, \cdots, y_{t-1}$ and $\beta_t \in \mathbb{R}^p$ and a non-random variance:

$$y_t = h_t(\beta_t, y_{t-1}, \cdots, y_0) + \epsilon_t, \tag{6.1}$$

  where $(\epsilon_t)_{t \in \mathbf{Z}}$ is a white noise sequence $\epsilon_t \stackrel{d}{=} \mathcal{N}(0, \Sigma_t)$, $\Sigma_t > 0$.

- the unobserved parameters β_t are stochastically modelled and satisfy a *linear* transition equation of the form:

$$\begin{aligned} \beta_{t+1} &= T_t \beta_t + \lambda_t \\ \beta_0 &= \alpha + \lambda_{-1} \end{aligned} \tag{6.2}$$

  with non-random transition matrices T_t and non-random initial condition α. The error process $(\lambda_t)_{t \in \mathbf{Z}}$ is white noise $\lambda_t \stackrel{d}{=} \mathcal{N}(0, Q_t)$, $Q_t > 0$ and independent of $(\epsilon_t)_{t \in \mathbf{Z}}$.

Note that Special zero-curvature models of time-varying autoregressive processes fit precisely in this theory, because the coefficient coordinates with respect to some chart follow a random walk.
Fahrmeir and Kaufmann design an algorithm to perform Gauss-Newton (G-N) iterations for finding the posterior mode, and they show the relation between the iteration steps and Extended Kalman Filtering and Smoothing (EKFS) (see Anderson and Moore (1979, chapter 8)) for models of the form (6.1, 6.2).
Also Jazwinski (1970), pp. 349-351, mentions procedures to perform posterior mode estimation for dynamical models, based on iterated Extended Kalman Filtering and Smoothing. These procedures are different from the (Gauss-)Newton method, proposed by Fahrmeir and Kaufmann.
We present their G-N algorithm and the comparison with EKFS in 6.1.5. However, we had to adapt their algorithm, because we want to apply their algorithm and analysis also to our more General zero-curvature models. We want to apply it to models where, with respect to some chart, coordinate coefficients follow a *smoothly integrated* random walk which has as a consequence that then the transition equation is not any more of the form (6.2). We are interested in these more general models, because in chapter four we were able to do asymptotics on a sequence of this type of models.
In the algorithm of Fahrmeir and Kaufmann, at every new Gauss-Newton iteration, a new observation is taken into account until all observations are processed. In order to do posterior mode estimation with a given number of observations it is, of course, necessary to repeat the Gauss-Newton iterations with all the observations until they convergence. Some small considerations about convergence are presented in 6.1.6.

In 6.1.7, we give some asymptotic results for different Riemannian metrics, related to observability of geodesic models. Uniform complete observability and uniform complete controllability play an important role in linear filter theory (see Jazwinski, 1970, chapter 7). We show that these concepts can also be used in the non-linear setting of geodesic models, here to bound the smallest eigenvalue of the gain-matrix in Fahrmeir/Kaufmann's algorithm away from zero.
In 6.1.8, the updating of the noise levels is treated. Special attention is paid to the "Gauss"-Newton gain-matrix including the determination of a lower bound for it. We also give an alternative to Newton iteration, inspired by Kitagawa and Gersch (1985).

6.1.1 (Approximate) expressions for the joint probability density

According to Bayes Rule, we can express the posterior density in the joint density and the marginal density of the observations:

$$p(q_0,\cdots,q_T|y_0,\cdots,y_T) = \frac{p(q_0,\cdots,q_T,y_0,\cdots,y_T)}{p(y_0,\cdots,y_T)}.$$

This shows that posterior mode estimation is the same as maximizing the joint density $p(q_0,\cdots,q_T,y_0,\cdots,y_T)$ with respect to $q_0,\cdots,q_T$. We shall give some (approximate) expressions for this joint density in this subsection.
According to assumptions 3, 4 and 5 in the beginning of this chapter, the transition equations of the General Model have the form:

$$q_0 = \exp_{q_{-1}}(\sum_{j=k-m+1}^{k} \omega_j \tau_j F_j); \tag{6.3}$$

$$q_{t+1} = \exp_{q_t}(\sum_{j=k-m+1}^{k} \theta_{t,j}\tau_j F_j) \text{ if } t \geq 0; \tag{6.4}$$

$$\omega \stackrel{d}{=} \mathcal{N}(0,Q^2); \tag{6.5}$$

$$\theta_{t+1} = A\theta_t + \lambda_t \text{ if } t \geq 0, \tag{6.6}$$

where $(\lambda_t)_{t\in\mathbb{Z}}$ is m-dimensional Gaussian unit variance white noise, matrix $A \in \mathbb{R}^{m\times m}$ is stable, and matrix $Q^2 \in \mathbb{R}^{m\times m}$ is positive definite. The number k is the dimension of the manifold M, and the vector fields F_j $(k-m+1 \leq j \leq k)$ of the direction space $\mathcal{L}$ are orthonormal.
We saw that, after choosing a chart $(\tilde{U}_\alpha, x_\alpha : \tilde{U}_\alpha \to V_\alpha \subset \mathbb{R}^k)$, we can extend the set of vector fields $F_{k-m+1},\cdots,F_k$ to an orthonormal base $(F_j)_{1\leq j\leq k}$ on $\tilde{U}_\alpha$, and express the coefficients $q_t = e^{-r_t}(x^n - a_{t,1}x^{n-1} + \cdots - a_{t,n})$ and vector fields F_j in these coordinates:

$$\psi = x_\alpha^{-1};\quad x_\alpha(q_t) = \kappa_t;\quad q_t = \psi(\kappa_t)\quad (q_t \in \tilde{U}_\alpha; \kappa_t \in V_\alpha);$$

$$F_j(q_t) = \sum_{i=1}^{k} (g_\alpha)_{ij}(\kappa_t)\frac{\partial}{\partial x_{\alpha\, i}},$$

where $g_\alpha : V_\alpha \to \mathbb{R}^{k\times k}$ is some positive definite matrix function . Linearizing equations (6.3, 6.4) w.r.t. the coordinates of this chart in combination with equations (6.5, 6.6) gives the following *approximate* model for the process $(y_t)_{t\geq 0}$:

$$y_t = \psi_1(\kappa_t)y_{t-1} + \cdots + \psi_n(\kappa_t)y_{t-n} + e^{\psi_{n+1}(\kappa_t)};\epsilon_t \tag{6.7}$$

$$\kappa_0 = \kappa_{-1} + g_\alpha(\kappa_{-1})C\omega \tag{6.8}$$

$$\theta_0 \stackrel{d}{=} \mathcal{N}(0, \sum_{k=0}^{\infty} A^k A'^k); \tag{6.9}$$

$$\begin{pmatrix} \kappa_{t+1} \\ \theta_{t+1} \end{pmatrix} = \begin{pmatrix} I & g_\alpha(\kappa_t)C \\ 0 & A \end{pmatrix} \begin{pmatrix} \kappa_t \\ \theta_t \end{pmatrix} + \begin{pmatrix} 0 \\ I \end{pmatrix} \lambda_t \text{ if } t \geq 0; \tag{6.10}$$

$$\text{where } C = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \\ \tau_{k-m+1} & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & \tau_k \end{pmatrix} \in \mathbb{R}^{k\times m}. \tag{6.11}$$

Here, $(\epsilon_t)_{t\in\mathbf{Z}}$ is Gaussian unit variance white noise independent of $(\lambda_t)_{t\in\mathbf{Z}}$ and independent of the initial conditions : the probability distributions of q_0 and $y_{-n}, \cdots, y_{-1}$. If the used general model is a zero-curvature model, then we can choose the parametrization ψ, such that equations (6.8, 6.10) are completely linear and there is no approximation in model (6.7-6.10).

Using this approximate General Model, it is now possible to calculate the joint probability distribution, where the pre-initial coefficient q_{-1} and pre-initial "observations" $y_{-n}, \cdots, y_{-1}$ are regarded fixed. If $m < k$, then this distribution is partly degenerate. The coordinates κ_t will have to satisfy $k - m$ fixed equations.

Now, denote minus the logarithm of the density of the non-degenerate part of this distribution for $T + 1$ observations and $T + 1$ coordinates of coefficients by J_T:

$$J_T = -\log p(\kappa_0, \cdots, \kappa_T, y_0, \cdots, y_T | \kappa_{-1}, y_{-n}, \cdots, y_{-1}).$$

If we have observations $y_0, \cdots, y_T$, then posterior mode estimation means that J_T is the criterion function that should be minimized with respect to $\kappa_0, \cdots, \kappa_{T+1}$. By Bayes Rule we have:

$$-\log p(\kappa_{t+1}, y_t | \kappa_{-1}, \cdots, \kappa_t, y_{-n}, \cdots, y_{t-1}) = p_t + l_t,$$

where

$$p_t = -\log p(\kappa_{t+1} | \kappa_{-1}, \cdots, \kappa_t, y_{-n}, \cdots, y_t)$$

(or the non-degenerate part of it) and

$$l_t = -\log p(y_t | \kappa_{-1}, \cdots, \kappa_t, y_{-n}, \cdots, y_{t-1}).$$

Also because of Bayes Rule, we can express the criterion function J_T as a sum of the terms l_t and p_t:

$$J_T = p_{-1} + \sum_{t=0}^{T-1}(l_t + p_t) + l_T.$$

First, we calculate the log-likelihood-terms l_t. We write

$$a(\kappa) := \begin{pmatrix} \psi_1(\kappa) \\ \vdots \\ \psi_n(\kappa) \end{pmatrix}; \quad r(\kappa) := \psi_{n+1}(\kappa); \quad \phi_t := \begin{pmatrix} y_{t-1} \\ \vdots \\ y_{t-n} \end{pmatrix}. \tag{6.12}$$

Conditional on past coefficients and data, the distribution of y_t is

$$y_t \stackrel{d}{=} \mathcal{N}(\phi_t' a(\kappa_t), \exp(2r(\kappa_t))),$$

hence, apart from the additive constant $\frac{1}{2}\log(2\pi)$,

$$l_t = r(\kappa_t) + \frac{1}{2} e^{-2r(\kappa_t)} (y_t - \phi_t' a(\kappa_t))^2. \tag{6.13}$$

This term consists of two components. One can be interpreted as an error of fitting y_t by $\phi_t' a(\kappa_t)$, weighted by the measurement noise $e^{r(\kappa_t)}$, the other being a penalty for allowing that measurement noise.
Next, we calculate the terms p_t. We introduce

$$P := \begin{pmatrix} 0 & \cdots & 0 & \frac{1}{\tau_{k-m+1}} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & \frac{1}{\tau_k} \end{pmatrix} \in \mathbb{R}^{m \times k} \tag{6.14}$$

$$h_t := g_\alpha^{-1}(\kappa_t)(\kappa_{t+1} - \kappa_t). \tag{6.15}$$

Then, $(h_{t\ 1}, \cdots, h_{t\ k-m}) = (0, \cdots, 0)$, which gives the degenerate part of the distribution of κ_{t+1}, given past coefficients and observations, and $(h_{t\ k-m+1}, \cdots, h_{t\ k})' = Ph_t = \theta_t$. Now, if $t > -1$, $\theta_t = A\theta_{t-1} + \lambda_{t-1}$, where λ_{t-1} is independent of the initial conditions $\kappa_{-1}, y_{-n}, \cdots, y_{-1}$, of $\cdots, \theta_s \cdots, \theta_{t-1}$ and the measurement noise $(\epsilon_s)_{s \geq 0}$. Hence λ_{t-1} is independent of the past $\kappa_{-1}, \cdots, \kappa_t, y_{-n}, \cdots, y_t$. Consequently, the distribution of Ph_t, conditional on this past, is normal with mean $A\theta_{t-1} = APh_{t-1}$ and unit variance. This gives the non-degenerate part of the conditional distribution of κ_{t+1}, hence, apart from the additive constant $\frac{m}{2}\log(2\pi)$,

$$p_t = \frac{1}{2}||Ph_t - APh_{t-1}||^2. \quad (t \geq 0); \tag{6.16}$$

$$p_{-1} = \frac{1}{2}||Q^{-1}Ph_{-1}||^2. \tag{6.17}$$

One can consider the term p_t as a penalty for changing the coefficient q_t into q_{t+1}. The change is not measured directly by the coordinate difference $\kappa_{t+1} - \kappa_t$, but by $h_t = g_\alpha(\kappa_t)^{-1}(\kappa_{t+1} - \kappa_t)$. The factor in front of the coordinate difference causes the coefficient change to be measured by the length of the vector represented by $\kappa_{t+1} - \kappa_t$ in the tangent space at the current coefficient $q_t = \psi(\kappa_t)$:

$$||h_t||^2 = \langle \kappa_{t+1} - \kappa_t, H_\alpha(\kappa_t)\ (\kappa_{t+1} - \kappa_t) \rangle = ||\sum_{i=1}^{k} (\kappa_{t+1} - \kappa_t)_i \frac{\partial}{\partial x_{\alpha\ i}}||_{q_t}^2.$$

Remark 6.1 If we use the asymptotic Fisher metric, then the term $||h_t||^2$ has an interesting form. It has the same form as used in the Wald statistic to test (for given κ_t) the restricted zero-hypothesis $\kappa_{t+1} = \kappa_t$ against the unrestricted alternative $\kappa_{t+1} \neq \kappa_t$ for a time-invariant stationary AR(n) process. The form is evaluated in $q_t = \psi(\kappa_t)$ (the "zero-hypothesis"). In zero-curvature models it doesn't matter where the form is evaluated; we get the same result in $\psi(\kappa_t)$ and in $\psi(\kappa_{t+1})$. □

6.1.2 Zero-curvature model and Newton algorithm

In this subsection, we start the discussion of the Newton method for posterior mode estimation by considering zero-curvature models. We have an exact expression for the joint probability density and linear equations (6.8, 6.10) on some parameterization for these models.

For a zero-curvature model on a k-dimensional manifold M with m-dimensional direction space $\mathcal{L}$, the point q_0 can be any point in the whole manifold M, for instance if we have a situation as discussed in chapter three, example 3.10. Subsequent coefficients q_t, $t > 0$, however, are lying on an m-dimensional, totally geodesic, integral manifold $N(q_0)$ of the distribution generated by the direction space $\mathcal{L}$. One can parametrize these manifolds by the first $k - m$ coordinates $\zeta = (\kappa_{0\,1}, \cdots, \kappa_{0\,k-m})$ of q_0. If we work conditional on this initial point q_0, and, consequently, assume the coordinates $\zeta = \kappa_{0\,1}, \cdots, \kappa_{0k-m}$ to be given, we can as well work with this manifold $N(q_0)$ instead of the manifold M. Nothing will change to the model. Hence, we shall suppose in this section that $k = m$ and that the matrix P is invertible. As the manifold $N(q_0)$ has curvature zero, it is possible to cover $N(q_0)$ with charts $(\tilde{U}_\beta \subset N(q_0), x_\beta : \tilde{U}_\beta \to V_\beta \subset \mathbb{R}^m)$ (that can be extended to charts on M), such that $\kappa \mapsto g^N_\beta(\kappa)$ is constant on V_β. We can even ensure that $g^N_\beta = I_m$. The transition equation is completely linear on these coordinates and the expressions for J_T of the preceding subsection are exact. The transition equations (6.8, 6.10) now have the form:

$$\kappa_0 = \kappa_{-1} + C\omega; \tag{6.18}$$

$$\theta_0 \stackrel{d}{=} \mathcal{N}(0, \sum_{k=0}^{\infty} A^k A'^k); \tag{6.19}$$

$$\begin{pmatrix} \kappa_{t+1} \\ \theta_{t+1} \end{pmatrix} = \begin{pmatrix} I & C \\ 0 & A \end{pmatrix} \begin{pmatrix} \kappa_t \\ \theta_t \end{pmatrix} + \begin{pmatrix} 0 \\ I \end{pmatrix} \lambda_t \text{ if } t \geq 0 \tag{6.20}$$

hence, the coordinates $(\kappa_t)_{t\geq 0}$ are modelled as a (smoothly integrated) random walk. We calculate the derivative of $J_T = \sum_{t=0}^{T} l_t + \sum_{t=-1}^{T-1} p_t$ with respect to $\kappa_0, \cdots, \kappa_T$. The transpose of this derivative will be denoted by:

$$f(\kappa_0, \cdots, \kappa_T) = \begin{pmatrix} f_0 \\ \vdots \\ f_T \end{pmatrix} = \begin{pmatrix} m_0 + c_0 \\ \vdots \\ m_T + c_T \end{pmatrix}. \tag{6.21}$$

Here, m is the derivative of $\sum_{t=0}^{T} l_t$ minus the conditional log-likelihood given the coefficients κ, and c is the derivative of $\sum_{t=-1}^{T-1} p_t$ minus the logarithm of the prior probability

density of $\kappa_0, \cdots, \kappa_T$.
The functions m_t are the score functions and are given by:

$$\begin{aligned} m_t &= \frac{\partial l_t'}{\partial \kappa}(\kappa_t) = (Dr'(\kappa_t))(1 - e^{-2r(\kappa_t)}(y_t - \phi_t' a(\kappa_t))^2) + \\ &\quad -(Da'(\kappa_t)\phi_t)e^{-2r(\kappa_t)}(y_t - \phi_t' a(\kappa_t)); \quad (0 \le t \le T) \end{aligned} \tag{6.22}$$

The terms c_t are given by:

$$\begin{aligned} c_0 &= -P'(I + A'A)P(\kappa_1 - \kappa_0) + P'(Q^{-2} + A'A)P(\kappa_0 - \kappa_{-1}) + \\ &\quad +P'AP(\kappa_0 - \kappa_{-1}) + \\ &\quad +P'A'P(\kappa_2 - 2\kappa_1 + \kappa_0); \end{aligned} \tag{6.23}$$

$$\begin{aligned} c_t &= -P'(I + A'A)P(\kappa_{t+1} - 2\kappa_t + \kappa_{t-1}) + \\ &\quad +P'AP(\kappa_t - 2\kappa_{t-1} + \kappa_{t-2}) + \\ &\quad +P'A'P(\kappa_{t+2} - 2\kappa_{t+1} + \kappa_t) \quad (1 \le t \le T-1); \\ c_{T-1} &= -P'P(\kappa_T - 2\kappa_{T-1} + \kappa_{T-2}) + P'A'AP(\kappa_{T-1} - \kappa_{T-2}) + \\ &\quad +P'AP(\kappa_{T-1} - 2\kappa_{T-2} + \kappa_{T-3}) + \\ &\quad +P'A'P(\kappa_{T-1} - \kappa_T); \end{aligned} \tag{6.24}$$

$$c_T = P'P(\kappa_T - \kappa_{T-1}) - P'AP(\kappa_{T-1} - \kappa_{T-2}). \tag{6.25}$$

Obviously, maximizing the posterior density on V_β or, equivalently, minimizing J_T, requires the computation of a zero of f. A common search routine to approximate a zero of a function is the following Newton algorithm:

$$\begin{aligned} &\text{choose } \hat{\kappa}^1 \quad \text{in a neighbourhood of zero } \kappa \\ &\text{update } \hat{\kappa}^p \text{ recursively as } \hat{\kappa}^{p+1} = \hat{\kappa}^p - (Df(\hat{\kappa}^p))^{-1} f(\hat{\kappa}^p) \end{aligned} \tag{6.26}$$

The sequence $(\hat{\kappa}^p)_{p \in \mathbb{N}}$ converges to a zero κ under suitable conditions. This zero will constitute a local maximum of the (conditional) joint density if the derivative of f in κ is positive definite. In modified versions of the Newton algorithm, the gain matrix $(Df(\hat{\kappa}^p))^{-1}$ may be replaced by other matrices, e.g. $(Df(\kappa))^{-1}$ or by some kind of conditional expectation of it. The derivative $U = Df(\hat{\kappa}^p)$ has the following block-five-diagonal structure:

$$U = \begin{pmatrix} U_{00} & U_{01} & U_{02} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ U_{10} & U_{11} & U_{12} & U_{13} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ U_{20} & U_{21} & U_{22} & U_{23} & U_{24} & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & U_{31} & U_{32} & U_{33} & U_{34} & U_{35} & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & U_{42} & U_{43} & U_{44} & U_{45} & U_{46} & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & U_{53} & U_{54} & U_{55} & U_{56} & U_{57} & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 & & \cdots \\ \vdots & \vdots & \vdots & \vdots & 0 & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & & & & & & & & & & \\ & & & & & & & \ddots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \vdots & & & & & & & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & \cdots & & & & & & & \ddots & \ddots & U_{TT} \end{pmatrix}. \tag{6.27}$$

Let us denote the derivative of the score functions by:

$$\tilde{M}_t(\hat{\kappa}_t^P) = \frac{\partial m_t}{\partial \kappa_t}(\hat{\kappa}_t^P).$$

Then, we can express the elements in the block matrix U as follows:

$$\begin{aligned}
U_{00} &= \tilde{M}_0 + P'(I + A'A)P + P'(Q^{-2} + A'A)P + P'AP + P'A'P; \\
U_{t\ t+1} &= U'_{t+1\ t} = -P'(I + A'A)P - 2P'A'P \quad \text{if } 0 \le t \le T-2; \\
U_{t\ t+2} &= U'_{t+2\ t} = P'A'P \quad \text{if } 0 \le t \le T-2; \\
U_{tt} &= \tilde{M}_t + 2P'(I + A'A)P + P'AP + P'A'P \quad \text{if } 1 \le t \le T-2; \\
U_{T-1\ T-1} &= \tilde{M}_{T-1} + P'(I + A'A)P + P'P + P'AP + P'A'P; \\
U_{T-1\ T} &= U'_{T\ T-1} = -P'P - P'A'P; \\
U_{T\ T} &= \tilde{M}_T + P'P.
\end{aligned} \tag{6.28}$$

Remark 6.2 If matrix A in the Model equations (6.3-6.6) is equal to zero, then the gain matrix U has only three non-zero diagonals. It was this situation that Fahrmeir and Kaufmann studied. The case $A = 0$ corresponds to the case that the process $(\theta_t)_{t\in\mathbf{Z}}$ is white noise, and $(\kappa_t)_{t\in\mathbf{Z}_+}$ is a random walk, i.e. it corresponds to Special Models. However, in chapter four we needed General Models, where $(\theta_t)_{t\in\mathbf{Z}}$ was an AR(1) process and $(\kappa_t)_{t\in\mathbf{Z}_+}$ was a smoothly integrated random walk. It was on a sequence of this type of models that we were able to define the stability and spectral density concept. We want to discuss the identification of these models too and, therefore, need to include the possibility that $A \neq 0$. We shall study the method of Fahrmeir and Kaufmann for this situation in subsection 6.1.4. □

6.1.3 Gauss-Newton iteration for zero-curvature models

The second derivative U of $J_T(\kappa)$ can also be written in another way. The coordinates κ are modelled as a smoothly integrated random walk, hence the prior distribution of $\kappa_0, \cdots, \kappa_T$ is Gaussian with positive definite variance $\mathcal{W}$. Minus the logarithm of the density of this prior is the term $\sum_{t=-1}^{T-1} p_t$ in the criterion $J_T(\kappa)$ (apart from a constant), hence its second derivative is equal to the inverse of $\mathcal{W}$. This yields the following expression for U:

$$\begin{aligned}
U &= D^2 J_T(\kappa) = D^2 \sum_{t=0}^{T} l_t + D^2 \sum_{t=-1}^{T-1} p_t = \\
&= \begin{pmatrix}
\tilde{M}_0 & 0 & & \cdots & \cdots & \cdots \\
0 & \tilde{M}_1 & 0 & \cdots & \cdots & \cdots \\
 & & \ddots & & & \\
\vdots & \vdots & & \ddots & & \vdots \\
 & & & & \ddots & \\
\cdots & \cdots & \cdots & & 0 & \tilde{M}_T
\end{pmatrix} + \mathcal{W}^{-1}.
\end{aligned} \tag{6.29}$$

The whole matrix $U(\kappa) = Df(\kappa)$ can be seen as a kind of Bayesian information matrix. One can consider $\tilde{M}_t(\kappa_t)$ to be the contribution of y_t to the information on κ_t. The *random information* $\tilde{M}_t$ can be decomposed in

$$\tilde{M}_t = MA_t + MB_t,$$

where

$$MA_t(\hat{\kappa}_t^p) = 2Dr'(\hat{\kappa}_t^p)Dr(\hat{\kappa}_t^p) + Da'(\hat{\kappa}_t^p)\frac{\phi_t\phi_t'}{e^{2r(\hat{\kappa}_t^p)}}Da(\hat{\kappa}_t^p), \tag{6.30}$$

and

$$\begin{aligned} MB_t \;=\; & D^2r'(\hat{\kappa}_t^p)(1 - e^{-2r(\hat{\kappa}_t^p)}(y_t - \phi_t'a(\hat{\kappa}_t^p))^2) + \\ & -2Dr'(\hat{\kappa}_t^p)Dr(\hat{\kappa}_t^p)(1 - e^{-2r(\hat{\kappa}_t^p)}(y_t - \phi_t'a(\hat{\kappa}_t^p))^2) + \\ & +Dr'(\hat{\kappa}_t^p)\phi_t'Da(\hat{\kappa}_t^p)e^{-2r(\hat{\kappa}_t^p)}(y_t - \phi_t'a(\hat{\kappa}_t^p)) + \\ & +Da'(\hat{\kappa}_t^p)\phi_tDr(\hat{\kappa}_t^p)e^{-2r(\hat{\kappa}_t^p)}(y_t - \phi_t'a(\hat{\kappa}_t^p)) + \\ & -D^2a'(\hat{\kappa}_t^p)\phi_te^{-2r(\hat{\kappa}_t^p)}(y_t - \phi_t'a(\hat{\kappa}_t^p)). \end{aligned} \tag{6.31}$$

Only the term MB_t depends on the current observation at time t, y_t. It depends on y_t in such a way that if one takes the expectation of $\tilde{M}_t(\kappa_t)$, conditional on coefficients $\kappa_0, \cdots, \kappa_t$ and past observations $y_{-n}, \cdots, y_{t-1}$,, this conditional expectation M_t is equal to $MA_t(\kappa_t)$: so the *conditional information* , defined by:

$$M_t = \boldsymbol{E}(\tilde{M}_t(\kappa_t)|\kappa_t, y_{-n}, \cdots, y_{t-1}), \tag{6.32}$$

equals

$$\boldsymbol{E}(MA_t(\kappa_t)|\kappa_t, y_{-n}, \cdots, y_{t-1}) = 2Dr'(\kappa_t)Dr(\kappa_t) + Da'(\kappa_t)\frac{\phi_t\phi_t'}{e^{2r(\kappa_t)}}Da(\kappa_t). \tag{6.33}$$

This conditional information is obviously non-negative definite.

We obtain Gauss-Newton-iteration if one replaces the random information $\tilde{M}_t$ in the gain-matrix U in the Newton iteration step by the conditional information M_t. It follows from equation (6.29) that the gain matrix is then strictly positive definite, hence the iteration can be carried out. Furthermore, if the noise-levels are taken fixed, the minimization is just a non-linear least squares problem, because the criterion can be written in the form:

$$\begin{aligned} J_T(\kappa) &= \frac{1}{2}n(\kappa)'n(\kappa), \text{ where} \\ n &= \begin{pmatrix} n_{-1} \\ \vdots \\ n_{2T} \end{pmatrix}; \\ n_{-1}(\kappa) &= Q^{-1}P; \\ n_{2k}(\kappa) &= e^{-r_k}(y_k - \phi_k'a(\kappa_k)) \quad (0 \le k \le T); \\ n_{2k-1}(\kappa) &= P(\kappa_k - \kappa_{k-1}) - AP(\kappa_{k-1} - \kappa_{k-2}) \quad (1 \le k \le T). \end{aligned} \tag{6.34}$$

It can easily be verified that $Dn'Dn = U$ and $Dn'n = f$.
There is another advantage of using Gauss-Newton iteration instead of Newton iteration. We have made the assumptions 1 and 2 in the beginning of this chapter. Hence, in equations (6.3-6.6), M is supposed to be an open part of $\frac{1}{\sigma}AR_n$ or U_n. Hence, $n \leq k \leq n+1$ and, if $k = n+1$, we can choose $F_{n+1} = \frac{1}{\mu}\frac{\partial}{\partial r}$, the log-noise-level vector field for some $\mu \in \mathbb{R}^+$, and matrix A has the form $A = \begin{pmatrix} \tilde{A} & 0 \\ 0 & a \end{pmatrix}$. For the chart $(\tilde{U}_\beta, x_\beta)$ we have: $g^N_\beta = I_m$, implying that $\frac{\partial}{\partial x_{\beta m}} = F_{n+1}$, hence $\rho := \kappa_m = \frac{1}{\mu}r$, and we can write:
$\psi_\beta(\kappa) = \psi_\beta(\begin{pmatrix} \tilde{\kappa} \\ \rho \end{pmatrix}) = e^{-\mu\rho}(x^n - a_1(\tilde{\kappa})x^{n-1} - \cdots - a_n(\tilde{\kappa}))$.
Now, the gain matrix U of the Gauss-Newton-iteration consists of two diagonal blocks (which each is a block-five-diagonal-matrix), the first only effecting the derivative of the log-joint density with respect to $\tilde{\kappa}_t$ $(0 \leq t \leq T)$, and the second only effecting the derivative with respect to ρ_t $(0 \leq t \leq T)$. The first half iteration step consists of updating the coefficients with given noise levels. In the second half step, we take the coefficients fixed and improve the estimates of the time-varying observation noise.

6.1.4 Identification of the zero-curvature model with given noise levels— decomposition of gain matrix

In this subsection, we study an LU decomposition of the gain-matrix U of the Gauss-Newton iteration step. This LU decomposition is the basis of the Fahrmeir/Kaufmann algorithm to carry out the Gauss-Newton step. We compare this decomposition with the decomposition of two other matrices: the matrix S of the conditional variance of the *states* given the observations for a *linearized* model, and the matrix $\mathcal{V}$ of the conditional variance of the *coordinates* κ_t , given the observations for the same linearized model. The decomposition of S can be obtained by means of Kalman Filtering and Smoothing with respect to the linearized model. If the components $\hat{\tilde{\kappa}}^p_t$ of the linearization point $\hat{\tilde{\kappa}}^p$ in the linearized model are well chosen, then Kalman Filtering and Smoothing with respect to the linearized model is just Extended Kalman Filtering and Smoothing with respect to the exact model. In this way, we (and Fahrmeir and Kaufmann) are able to relate EKFS to G-N iteration.

Exact model.
If the noise levels σ_t are given, the zero-curvature General Model (6.7-6.10) reduces to the following form:

$$\begin{aligned} \tilde{\kappa}_{t+1} &= \tilde{\kappa}_t + \tilde{C}\tilde{\theta}_t \\ \tilde{\theta}_{t+1} &= \tilde{A}\tilde{\theta}_t + \tilde{\lambda}_t \\ y_t &= \phi_t' a(\tilde{\kappa}_t) + \sigma_t \epsilon_t, \end{aligned} \tag{6.35}$$

where $\begin{pmatrix} \tilde{\lambda}_t \\ \epsilon_t \end{pmatrix} \overset{i.i.d.}{=} \mathcal{N}(0, I_m)$, matrix $\tilde{A} \in \mathbb{R}^{m-1 \times m-1}$ is stable, and matrix $\tilde{C} \in \mathbb{R}^{m-1 \times m-1}$ has inverse $\tilde{P}$.

This is in fact the zero-curvature model for the *projected* instantaneous associated polynomials on AR_n, see chapter five, section 5.3. □

We study the first-half-Gauss-Newton-iteration step from point $\hat{\kappa}^p = (\hat{\kappa}_0^p, \cdots, \hat{\kappa}_T^p)$ to the following point $\hat{\kappa}^{p+1}$. We compare it with *Kalman filtering and smoothing* for a model which differs from (6.35) in the fact that the measurement equation $a(\tilde{\kappa}_t)$ is linearized and replaced by:

$$a(\tilde{\kappa}_t) = a(\hat{\tilde{\kappa}}_t^p) + Da(\hat{\tilde{\kappa}}_t^p)(\tilde{\kappa}_t - \hat{\tilde{\kappa}}_t^p).$$

Linearized model.
Introduce the $2(m-1)$-dimensional states

$$x_t = \begin{pmatrix} x_t^1 \\ x_t^2 \end{pmatrix} = \begin{pmatrix} \tilde{\kappa}_t \\ \tilde{C}\tilde{\theta}_{t-1} \end{pmatrix},$$

the filtered observations

$$z_t = y_t - \phi_t' a(\hat{\tilde{\kappa}}_t^p) + \phi_t' Da(\hat{\tilde{\kappa}}_t^p)\hat{\tilde{\kappa}}_t^p,$$

and the system matrices

$$\mathcal{A} = \begin{pmatrix} I & \tilde{C}\tilde{A}\tilde{C}^{-1} \\ 0 & \tilde{C}\tilde{A}\tilde{C}^{-1} \end{pmatrix}; \quad \mathcal{B} = \begin{pmatrix} \tilde{C} \\ \tilde{C} \end{pmatrix}; \quad C_t = \phi_t' Da(\hat{\tilde{\kappa}}_t^p)(I \;\; 0).$$

Then, model (6.35) with the linearized measurement equation reads as follows:

$$\begin{aligned} x_{t+1} &= \mathcal{A}x_t + \mathcal{B}\lambda_{t-1}, \\ z_t &= C_t x_t + \sigma_t \epsilon_t; \\ x_0 &\stackrel{d}{=} \mathcal{N}\left(\begin{pmatrix} \tilde{\kappa}_{-1} \\ 0 \end{pmatrix}, \begin{pmatrix} \tilde{C}Q^2\tilde{C}' & \tilde{C}Q^2\tilde{C}' \\ \tilde{C}Q^2\tilde{C}' & \tilde{C}Q^2\tilde{C}' \end{pmatrix}\right). \end{aligned} \qquad \square \tag{6.36}$$

The posterior density $p(x_0, \cdots, x_T | x_{-1}, z_0, \cdots, z_T)$ of this *linearized* model is Gaussian, hence, the posterior density $p(\tilde{\kappa}_0, \cdots, \tilde{\kappa}_T | \kappa_{-1}, y_{-n}, \cdots, y_T)$ is based on this model (6.36). The derivative of minus the logarithm of the latter posterior density w.r.t. $\tilde{\kappa}$ in $\hat{\kappa}^p$ is just given by (the first part of) $f(\hat{\kappa}^p)$. The second derivative of minus the logarithm of the latter posterior density is equal to the Gauss-Newton gain matrix U in the point $\hat{\kappa}^p$, and is hence positive definite. Its inverse, clearly, coincides with the variance $\mathcal{V}$ of the posterior distribution w.r.t. $\tilde{\kappa}$ for the linearized model, and is therefore also positive definite. However, the variance S of the posterior distribution of the states $x_0, \cdots, x_T$ can't be strictly positive definite. This follows from the definition of the states x_t:

$$x_t = \begin{pmatrix} \tilde{\kappa}_t \\ \tilde{\kappa}_t - \tilde{\kappa}_{t-1} \end{pmatrix},$$

hence, we have the noise-free equations

$$(x_t)_1 - (x_{t-1})_1 = (x_t)_2 \tag{6.37}$$

for all t, $0 \leq t \leq T$, so the dimension of the kernel of the variance S (seen as matrix of a linear map) is at least half its whole number of columns. Because the variance of $\mathcal{V}$ is positive definite, the rank of the variance S is exactly half its number of columns. This is also clear from the following lemma that gives a *decompositon for the matrix* S.

Lemma 6.3 . The variance S of the posterior distribution of the states $x_0, \cdots, x_T$, according to model (6.36), has the following decompositon:
There are matrices L and D of the form

$$L = \begin{pmatrix} I & -F_0 & & 0 & \cdots & 0 \\ 0 & I & -F_1 & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & I & -F_{T-1} \\ 0 & 0 & 0 & \cdots & 0 & I \end{pmatrix}; \quad D = \begin{pmatrix} D_0 & 0 & & \cdots & \cdots & \cdots \\ 0 & D_1 & 0 & \cdots & \cdots & \cdots \\ & & \ddots & & & \\ \vdots & \vdots & & \ddots & & \vdots \\ & & & & \ddots & \\ \cdots & \cdots & \cdots & & 0 & D_T \end{pmatrix} \tag{6.38}$$

such that $LSL' = D$, and such that $D_0 = 0$; D_T has full rank and the rank of D_t is half the rank of D_T for all t, $1 \leq t \leq T-1$.
The matrices F_t are the Kalman fixed interval smoothing gain matrices, and D_T is the conditional variance of the state x_T, given all observations, according to the linearized model (6.36).

Proof: The proof is based on well-known relations from Kalman Filtering and Smoothing theory.
We denote the conditional expectation of x_t, given $z_0, \cdots, z_k$ (or, equivalently, $y_0, \cdots, y_k$) by $x_{t|k}$ and its conditional variance by $\Sigma_{t|k}$. These variances satisfy the well-known rules

$$\Sigma_{t+1|t} = \mathcal{A}\Sigma_{t|t}\mathcal{A}' + \mathcal{B}\mathcal{B}' \tag{6.39}$$

$$\Sigma_{t|t} = (I + \Sigma_{t|t-1}\begin{pmatrix} M_t & 0 \\ 0 & 0 \end{pmatrix})^{-1}\Sigma_{t|t-1} \text{ where} \tag{6.40}$$

$$\begin{pmatrix} M_t & 0 \\ 0 & 0 \end{pmatrix} = \frac{C_t'C_t}{\sigma_t^2}. \tag{6.41}$$

Note that the matrix $\begin{pmatrix} \mathcal{B} & \mathcal{A}\mathcal{B} \end{pmatrix}$ has full rank, hence the linear system (6.36) is controllable.
Although $\Sigma_{0|0}$ is singular (but half-rank), we have that $\Sigma_{1|0}$ is non-singular. This can be shown as follows: $\Sigma_{0|-1} = \mathcal{B}Q^2\mathcal{B}'$ and $\Sigma_{1|0}x = 0$ implies that both $\mathcal{B}'x = 0$ and $\Sigma_{0|0}\mathcal{A}'x = 0$. This yields $\Sigma_{0|-1}\mathcal{A}'x = 0$, hence $\mathcal{B}'\mathcal{A}'x = 0$. As the rank of matrix $\begin{pmatrix} \mathcal{B} & \mathcal{A}\mathcal{B} \end{pmatrix}$ is full, we conclude that $x = 0$.
The linear system (6.36) is controllable, and, consequently, all $\Sigma_{t|t}$ and $\Sigma_{t+1|t}$ are positive definite for all $t \geq 1$ [1]. This enables us to define the (smoothing gain) operators F_t for all t, $0 \leq t \leq T-1$ by:

$$F_t = \Sigma_{t|t}\mathcal{A}'\Sigma_{t+1|t}^{-1}. \tag{6.42}$$

[1] See e.g. Jazwinski, lemma 7.3. Note that in the terminology of that book, the system is uniformly completely controllable with $N = 1$.

Then, we have the following well known property for the conditional covariances:

$$\begin{aligned} \text{Let } S_{t\ t+k} &:= \boldsymbol{E}((x_t - x_{t|T})(x_{t+k} - x_{t+k|T})'|z_0, \cdots, z_T) \qquad (0 \le t < t+k \le T) \\ \text{then, } S_{t\ t+k} &= F_t \cdots F_{t+k-1} \Sigma_{t+k|T}. \end{aligned} \tag{6.43}$$

This relation shows the first claim of the lemma, i.e. we can decompose the matrix of the variance S as follows:

$$\text{Let } L_0 = \begin{pmatrix} I & -F_0 & 0 & \cdots & 0 \\ 0 & I & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & I \end{pmatrix}, \cdots, \quad L_{T-1} = \begin{pmatrix} I & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & I & 0 & 0 \\ 0 & \cdots & 0 & I & -F_{T-1} \\ 0 & \cdots & 0 & 0 & I \end{pmatrix}$$

then, with $L = L_{T-1} \cdots L_0$, we have:

$$LSL' = \begin{pmatrix} S_{00} - F_0 S_{11} F_0' & 0 & \cdots & 0 & 0 \\ 0 & S_{11} - F_1 S_{22} F_1' & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & S_{T-1\ T-1} - F_{T-1} S_{TT} F_{T-1}' & 0 \\ 0 & 0 & \cdots & 0 & S_{TT} \end{pmatrix}. \tag{6.44}$$

Furthermore, we can rewrite every diagonal element $D_t := \Sigma_{t|T} - F_t \Sigma_{t+1|T} F_t'$ on the right hand side of this equation (6.44)from which by using the well known fact

$$D_t = \Sigma_{t|T} - F_t \Sigma_{t+1|T} F_t' = \Sigma_{t|t} - F_t \Sigma_{t+1|t} F_t' \quad (0 \le t \le T-1). \tag{6.45}$$

Now, we are going to compute the ranks of the matrices D_t, $(0 \le t \le T)$. As proven above, $D_T = \Sigma_{T|T}$ has full rank. In order to calculate the rank of the other D_t, $0 \le t \le T-1$, we note that we can decompose the matrix

$$\tilde{S}_t = \begin{pmatrix} \Sigma_{t|t} & \Sigma_{t|t} \mathcal{A}' \\ \mathcal{A} \Sigma_{t|t} & \Sigma_{t+1|t} \end{pmatrix}$$

in two ways: firstly,

$$\begin{pmatrix} I & 0 \\ -\mathcal{A} & I \end{pmatrix} \tilde{S}_t \begin{pmatrix} I & -\mathcal{A}' \\ 0 & I \end{pmatrix} = \begin{pmatrix} \Sigma_{t|t} & 0 \\ 0 & \mathcal{B}\mathcal{B}' \end{pmatrix};$$

and secondly,

$$\begin{pmatrix} I & -F_t \\ 0 & I \end{pmatrix} \tilde{S}_t \begin{pmatrix} I & 0 \\ -F_t' & I \end{pmatrix} = \begin{pmatrix} D_t & 0 \\ 0 & \Sigma_{t+1|t} \end{pmatrix}.$$

This shows that for $0 < t \le T-1$,

$$\text{rank } D_t = \text{rank } \Sigma_{t|t} + \text{rank } \mathcal{B}\mathcal{B}' - \text{rank } \Sigma_{t+1|t} = \text{rank } \mathcal{B}\mathcal{B}',$$

So every D_t has half rank for $0 < t \le T-1$, whereas for $t = 0$ we have:

$$\text{rank } D_0 = \text{rank } \Sigma_{0|0} + \text{rank } \mathcal{B}\mathcal{B}' - \text{rank } \Sigma_{1|0} = 0.$$

This completes the proof of lemma 6.3. □

In order to be able to compare the Gauss-Newton method with the Extended Kalman Filter plus Smoother, we first compare the decomposition of the Gauss-Newton gain matrix U and its inverse $\mathcal{V}$ with the decomposition of S.

Because the block-five-diagonal matrix U is positive definite, one can apply Gaussian elimination and decompose U as follows:

$$\text{Let } \Lambda = \begin{pmatrix} I & -B_0 & C_0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & I & -B_1 & C_1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & I & -B_2 & C_2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & -B_{T-3} & C_{T-3} & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & I & -B_{T-2} & C_{T-2} \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & I & -B_{T-1} \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & I \end{pmatrix}$$

$$\text{then } U = \Lambda' \begin{pmatrix} \Delta_0 & 0 & \cdots & 0 \\ 0 & \Delta_1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \Delta_T \end{pmatrix} \Lambda. \tag{6.46}$$

The matrices B_k, C_k, Δ_k are implicitly defined by the following equations from which these matrices can be solved forward recursively:

$$\begin{aligned} \Delta_0 &= U_{00} \\ -\Delta_0 B_0 &= U_{01} \\ B_0' \Delta_0 B_0 + \Delta_1 &= U_{11} \\ \Delta_k C_k &= U_{k\ k+2} \\ -B_k' \Delta_k C_k - \Delta_{k+1} B_{k+1} &= U_{k+1\ k+2} \\ C_k' \Delta_k C_k + B_{k+1}' \Delta_{k+1} B_{k+1} + \Delta_{k+2} &= U_{k+2\ k+2}. \end{aligned} \tag{6.47}$$

The variance $\mathcal{V}$ is, according to the linearized model (6.36), the inverse of the Gauss Newton matrix. Hence, we have:

$$\begin{pmatrix} \Delta_0^{-1} & 0 & \cdots & 0 \\ 0 & \Delta_1^{-1} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \Delta_T^{-1} \end{pmatrix} = \Lambda \mathcal{V} \Lambda'. \tag{6.48}$$

The matrix $\mathcal{V}$ was already decomposed as part of the matrix S. Hence, we can relate B_t and C_t ($0 \leq t \leq T-1$ to the *Kalman smoothing gain matrices* F_t ($0 \leq t \leq T-1$) and the matrices D_t to Δ_t^{-1}.

Lemma 6.4 1. The smoothing gain matrices F_t, $0 \leq t \leq T-1$ can be expressed in the matrices B_t and C_t as follows:

$$\begin{aligned} F_t &= \begin{pmatrix} I & -I \\ C_{t-1} - B_{t-1} + I & B_{t-1} - I \end{pmatrix} = \\ &\quad \begin{pmatrix} I & -I \\ I & 0 \end{pmatrix}^{-1} \begin{pmatrix} B_{t-1} & -C_{t-1} \\ I & 0 \end{pmatrix} \begin{pmatrix} I & -I \\ I & 0 \end{pmatrix} \\ &\quad (1 \leq t \leq T-2) \end{aligned} \tag{6.49}$$

$$F_0 = \begin{pmatrix} I & -I \\ I & -I \end{pmatrix}, \text{ and } F_{T-1} = \begin{pmatrix} B_{T-1} & 0 \\ C_{T-2} - B_{T-2} + I & B_{T-2} - I \end{pmatrix}. \tag{6.50}$$

2. The matrices D_t can be expressed in the matrices Δ_t^{-1} as follows:

$$\begin{aligned} D_t &= \begin{pmatrix} 0 & 0 \\ 0 & \Delta_{t-1}^{-1} \end{pmatrix} \quad (1 \leq t \leq T-2); \\ D_T &= \begin{pmatrix} \Delta_T^{-1} & 0 \\ 0 & \Delta_{T-1}^{-1} \end{pmatrix}; \quad D_0 = 0. \end{aligned} \tag{6.51}$$

3. The matrices Δ_t^{-1} satisfy the (in)equalities

$$\Delta_T^{-1} = \mathcal{V}_{TT}, \text{ and} \tag{6.52}$$

$$\Delta_t^{-1} < \mathcal{V}_{tt} \quad (0 \leq t < T). \tag{6.53}$$

Point 3 can be interpreted as follows. If $\hat{\hat{\kappa}}^p$ is close to the maximizing κ of the posterior of the original model, we can interprete Δ_T^{-1} as the approximate error covariance of $\tilde{\kappa}_T$, given all data $y_0, \cdots, y_T$.

Proof of Lemma: If we define $\mathcal{V}_{jk} = 0$ for all $j > T$ or $k > T$, then equation (6.48) yields the relations:

$$\begin{aligned} \Delta_k^{-1} &= \mathcal{V}_{kk} - \mathcal{V}_{k\,k+1}B_k' + \mathcal{V}_{k\,k+2}C_k' + \\ &\quad -B_k\mathcal{V}_{k+1\,k} + B_k\mathcal{V}_{k+1\,k+1}B_k' - B_k\mathcal{V}_{k+1\,k+2}C_k' + \\ &\quad +C_k\mathcal{V}_{k+2\,k} - C_k\mathcal{V}_{k+2\,k+1}B_k' + C_k\mathcal{V}_{k+2\,k+2}C_k' = \\ &= \boldsymbol{Var}(\kappa_k - B_k\kappa_{k+1} + C_k\kappa_{k+2}|z_0, \cdots, z_T); \\ 0 &= \mathcal{V}_{mk} - \mathcal{V}_{m\,k+1}B_k' + \mathcal{V}_{m\,k+2}C_k' + \\ &\quad -B_m\mathcal{V}_{m+1\,k} + B_m\mathcal{V}_{m+1\,k+1}B_k' - B_m\mathcal{V}_{m+1\,k+2}C_k' + \\ &\quad +C_m\mathcal{V}_{m+2\,k} - C_m\mathcal{V}_{m+2\,k+1}B_k' + C_m\mathcal{V}_{m+2\,k+2}C_k' = \\ &= \boldsymbol{Cov}(\kappa_m - B_m\kappa_{m+1} + C_m\kappa_{m+2}, \kappa_k - B_k\kappa_{k+1} + C_k\kappa_{k+2}|z_0, \cdots, z_T) \\ &\quad \text{for all } 0 \leq k, m \leq T. \end{aligned} \tag{6.54}$$

This shows that, in accordance with the linearized model (6.36), the sequence

$$\kappa_T, \quad \kappa_{T-1} - B_{T-1}\kappa_t, \quad \kappa_{T-2} - B_{T-2}\kappa_{T-1} + C_{T-2}\kappa_T, \quad \kappa_{T-3} - B_{T-3}\kappa_{T-2} + C_{T-3}\kappa_{T-1}, \cdots$$

consists of mutually conditionally independent variables. Hence,

$$\boldsymbol{Var}(\kappa_t - B_t\kappa_{t+1} + C_t\kappa_{t+2}|z_0,\cdots,z_T) \leq \boldsymbol{Var}(\kappa_t|z_0,\cdots,z_T) = \mathcal{V}_{tt}.$$

This immediately proves the third point of the Lemma, and in combination with the following three facts:

- $x_t = \begin{pmatrix} \kappa_t \\ \kappa_t - \kappa_{t-1} \end{pmatrix}$;
- $D_t = \boldsymbol{Var}(x_t - F_t x_{t+1}|z_0,\cdots,z_T)$, hence the variables $x_T, x_{T-1} - F_{T-1}x_T, \cdots, x_t - F_t x_{t+1}$ are conditionally independent;
- rank D_t=rank Δ_t if $1 \leq t \leq T-1$,

it also proves the first two points. □

6.1.5 Identification of the zero-curvature model with given noise levels— Fahrmeir/Kaufmann's algorithm

Now, we return to the Gauss-Newton iteration. We are now able to calculate the iteration step $\tilde{\kappa}^{p+1} = \tilde{\kappa}^p - U^{-1}f$. Following Fahrmeir and Kaufmann, we compute the step $U^{-1}f$ as follows:

$$\text{Let } \delta = U^{-1}f, \text{ or equivalently, } \Lambda'(\begin{pmatrix} \Delta_0 & 0 & \cdots & 0 \\ 0 & \Delta_1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \Delta_T \end{pmatrix} \Lambda\delta) = f.$$

$$\text{First, we solve } \Lambda'\xi = f, \text{ and next } \begin{pmatrix} \Delta_0 & 0 & \cdots & 0 \\ 0 & \Delta_1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \Delta_T \end{pmatrix} \Lambda\delta = \xi. \tag{6.55}$$

Solving the first equation, is just performing the forward recursion

$$\begin{aligned} \xi_0 &= f_0 \\ \xi_1 &= B_0'\xi_0 + f_1 \\ \xi_{t+1} &= B_t'\xi_t - C_{t-1}'\xi_{t-1} + f_{t+1} \\ &\quad \text{for } 1 \leq t \leq T-1, \end{aligned} \tag{6.56}$$

while the second equation (6.55) can be solved backward recursively:

$$\begin{aligned} \delta_T &= \Delta_T^{-1}\xi_T \\ \delta_{T-1} &= B_{T-1}\delta_T + \Delta_{T-1}^{-1}\xi_{T-1} \\ \delta_{t-1} &= B_{t-1}\delta_t - C_{t-1}\delta_{t+1} + \Delta_{t-1}^{-1}\xi_{t-1} \\ &\quad \text{for } 1 \leq t \leq T-1. \end{aligned} \tag{6.57}$$

After carrying out one Newton iteration step, based on the $T+1$ observations $y_0, \cdots, y_T$, the data analist has two options. Firstly, he can continue the procedure and perform another step in order to improve the approximation. Secondly, he can consider a new observation y_{T+1}, and regard the found estimate $\hat{\tilde{\kappa}}^{p+1}$ as a starting point for the Newton search of the maximizing $\tilde{\kappa}$ of the posterior density, based on $T+2$ observations. Of course, he has to add one new $\hat{\tilde{\kappa}}^{p+1}_{T+1}$ to initialize the algorithm, but this can be accomplished by the prediction rule

$$x_{T+1|T} = \mathcal{A} x_{T|T}$$

in combination with

$$(x_{T|T})_2 = (x_{T|T})_1 - (x_{T-1|T})_1$$

yielding:

$$\hat{\tilde{\kappa}}^{p+1}_{T+1} = \hat{\tilde{\kappa}}^{p+1}_{T} + \tilde{C}\tilde{A}\tilde{C}^{-1}(\hat{\tilde{\kappa}}^{p+1}_{T} - \hat{\tilde{\kappa}}^{p+1}_{T-1}). \tag{6.58}$$

In this way, an initialization point for the Newton search with T observations can be found by just performing one iteration step of the search for the first t observations, for all $t \leq T$. The method of Fahrmeir and Kaufmann becomes:

1. *Initialization:*

$$\hat{\tilde{\kappa}}^{0}_{-1} = \tilde{\kappa}_{-1}; \quad (\hat{\tilde{\kappa}}^{0}_{k} = 0 \quad \text{for } k < -1);$$

for $t = 0, \cdots, T$;

2. *Prediction step:*

$$\hat{\tilde{\kappa}}^{t}_{t} = \hat{\tilde{\kappa}}^{t}_{t-1} + \tilde{C}\tilde{A}\tilde{C}^{-1}(\hat{\tilde{\kappa}}^{t}_{t-1} - \hat{\tilde{\kappa}}^{t}_{t-2});$$

3. *Forward recursion, for $s = 0, \cdots, t$;*
compute $m_s, c_s, f_s = m_s - c_s, M_s, U_{ss}, \Delta_s, B_s, C_s, \xi_s$, using equations (6.22) for m_s;

$$m_s = -(Da'(\hat{\tilde{\kappa}}^{t}_{s})\phi_s e^{-2r_s})(y_s - \phi_s' a(\hat{\tilde{\kappa}}^{t}_{s}));$$

(6.23-6.25) for c_s; (6.33) for M_s;

$$M_s = Da'(\hat{\tilde{\kappa}}^{t}_{s})\frac{\phi_s \phi_s'}{e^{2r_s}} Da(\hat{\tilde{\kappa}}^{t}_{s});$$

(6.28) for U_{ss}.
For Δ_s, B_s, C_s, we can use (6.47) or we can use (6.49) and (6.51), and the covariance recursion for $\Sigma_{t|t}$, $\Sigma_{t+1|t}$, F_t ,D_t from equations (6.39), (6.40), (6.42), (6.45);
For ξ_s we use (6.56).

4. *Filter correction:*
First equation of the backward recursion (6.57)

$$\delta_t = \Delta_t^{-1}\xi_t$$

combined with

$$\hat{\hat{\kappa}}_t^{t+1} = \hat{\hat{\kappa}}_t^t - \delta_t.$$

Note that, according to the linearized model (6.36), $\Delta_t^{-1} = \boldsymbol{Var}(\kappa_t | z_0, \cdots, z_t)$, see equation (6.52).

5. *Smoother corrections. For* $s = t-1, \cdots, s = 0$.
The other equations of the backward recursion (6.57) and

$$\hat{\hat{\kappa}}_s^{t+1} = \hat{\hat{\kappa}}_s^t - \delta_s.$$

(*Remark:* Having stored $\Sigma_{s|s}$ $(0 \leq s \leq t-1)$, one can wait until this step before computing Δ_s^{-1} from D_s by (6.51), and D_s from $\Sigma_{s|s}$ by (6.39); (6.42); (6.45).)

For a linear model like (6.36), performing steps 2, 3, 4, 5 at time $t = t_1$ yields immediately the mode or the conditional expectation of $\kappa_0, \cdots, \kappa_{t_1}$ given $y_0, \cdots, y_{t_1}$, because f is linear and U is exactly the derivative. If we, subsequently, carry out the part of the method based on $y_0, \cdots, y_{t_1+1}$, then we obtain $\xi_0 = 0, \cdots, \xi_{t_1} = 0; \xi_{t_1+1} = f_{t_1+1}$.
From the definition of the terms c_{t_1+1} (6.25) and the definition of $\hat{\hat{\kappa}}_{t_1+1}^{t_1+1}$ in the prediction step, it follows that $f_{t_1+1} = m_{t_1+1}$ in the part of the method at time $t_1 + 1$. Hence, with the help of the first equation of the backward recursion (6.57) and equation (6.52), we obtain that

$$\hat{\hat{\kappa}}_{t_1+1}^{t_1+2} = \hat{\hat{\kappa}}_{t_1+1}^{t_1+1} - \mathcal{V}_{t_1+1\ t_1+1} m_{t_1+1}. \tag{6.59}$$

This is the same filtering update equation as we get if we apply the Kalman Filter to model (6.36). The whole backward recursion (6.57) coincides with the Kalman Smoothing algorithm for this model, because the resulting $\hat{\hat{\kappa}}_k^{t_1+2} = \hat{\hat{\kappa}}_k^{t_1+1} - \delta_k$ will give us immediately the conditional expectation of κ_k, given $y_0, \cdots, y_{t+1}$. Hence, step 5 is superfluous for the linear model for times $t < T$. It is enough to perform this step only once, when we base us on all the observations $y_0, \cdots, y_T$, because we are merely interested in the conditional expectations of the κ_t, given all the data.

Recapitulating, for the linear model, we see that Fahmeir's method immediately gives us the desired mode. The only steps that have to be carried out are prediction step 2 and the filter correction step 4, or equivalently, equation (6.59), and these are the same as the prediction and filter correction step in the Kalman Filter. From the last part of the method, based on all data, also step 5 has to be carried out, and this recursion is just Kalman Smoothing. Hence, the method we discussed above for the linear model reduces to Kalman Filtering and Smoothing.
Now, we compare the proposed method of Fahrmeir and Kaufmann with the ordinary Extended Kalman Filter (EKF) for the non-linear zero-curvature model. The algorithm is just Kalman Filtering on the *linearized* model (6.36), where the components $\hat{\hat{\kappa}}_t^P$ of the point of linearization $\hat{\hat{\kappa}}^P$ are recursively determined by the Kalman Filter itself. Hence, EKF is the Fahrmeir/Kaufmann's procedure, in which in particular steps 3 and 5 are not carried out. The calculation of the ξ_k's is omitted as well as the backward recursion (smoothing) at every instant t after which all residuals or scores m_k, the informations M_k, the gain matrices B_k, C_k and Δ_k $(k \leq t+1)$ are recomputed.

6.1.6 Repetitions of Fahrmeir/Kaufmann's method based on the same set of data

Fahrmeir/Kaufmann's procedure was just the method that produced a starting point $\hat{\tilde{\kappa}}^T$ for the Gauss-Newton iteration step, based on *all* data $y_0, \cdots, y_T$. If the procedure is completed, already more work is done than in applying the Extended Kalman Filter. But one can proceed.

For the time being, we suppose that the fixed noise-levels are already the ones corresponding to a minimum of the criterion if such a minimum exists in the region (V_α) we consider. Then, the second part of the G-N iteration step is not necessary anymore. Now, we can repeat steps 3, 4, 5 of the method. This is just another Gauss-Newton iteration step, based on the same data. We obtain a sequence of estimates $(\hat{\tilde{\kappa}}^p)_{p\in\mathbf{N}}$. Sufficient conditions for the convergence of the Gauss-Newton estimates to a minimum $\tilde{\kappa}^* \in \tilde{V}_\alpha$ of $J_T(\kappa)$ (if it exists) are given in Theorem 10.2.1 in the book of Dennis and Schnabel (1983). We use the notations of equations (6.34). The most critical condition of the theorem is that the *smallest eigenvalue of the gain matrix* $U(\tilde{\kappa}^*) = Dn(\tilde{\kappa}^*)'Dn(\tilde{\kappa}^*)$ should be larger than the matrix norm of $D^2n(\tilde{\kappa}^*)'n(\tilde{\kappa}^*)$. If this condition is satisfied, the theorem guarantees the existence of a neighbourhood of $\tilde{\kappa}^*$, such that any G-N sequence starting in this neighbourhood converges to $\tilde{\kappa}^*$.

We have assumed that the ρ_t's are already minimizing, hence we are only working with J_T as a function of $\tilde{\kappa}$. We can drop all terms with Dr or D^2r in the expressions MA_t and MB_t for the derivative Df in equations (6.33, 6.31). Consequently, $D^2n(\tilde{\kappa}^*)'n(\tilde{\kappa}^*)$ is just the (symmetric) part of the derivative $Df(\tilde{\kappa}^*)$ that was dropped in the gain matrix $U(\tilde{\kappa}^*)$. The matrix norm of $D^2n(\tilde{\kappa}^*)'n(\tilde{\kappa}^*)$ is just the length of its largest eigenvalue. Hence, $Df(\tilde{\kappa}^*)$ is positive definite, and $\tilde{\kappa}^*$ constitutes a local minimum of the criterion if the condition of the theorem is met.

If condition of theorem 10.2.1 in the book of Dennis and Schnabel is not satisfied, but there exists a minimizing $\tilde{\kappa}^*$ in $\tilde{V}_\alpha$, then there are various alternatives. For instance, one could use the exact derivative Df instead of U. For a minimizing $\tilde{\kappa}^*$, the derivative $Df(\tilde{\kappa}^*)$ is non-negative definite. If it is even positive definite, then $Df(\kappa)$ will be positive definite in a neigbourhood of $\tilde{\kappa}^*$. In that neighbourhood Fahrmeir/Kaufmann's method can be used to calculate the minimum, because the algorithm (step 1,2,3,4,5) works also with $U = Df(\hat{\tilde{\kappa}}^p)$ if this matrix is positive definite. The method yields a sequence of estimates, convergent to $\tilde{\kappa}^*$, if we are close enough to $\tilde{\kappa}^*$, according to Theorem 5.2.1, Dennis and Schnabel. We shall show below that it is possible to compute (a good approximation of) $Df(\hat{\tilde{\kappa}}^p)$ even if an analytic expression of the parametrization $: \tilde{V}_\alpha \to N(q_0); \tilde{\kappa} \mapsto a(\tilde{\kappa})$ is not known, and only the metric itself is available.

Other alternatives to strict Gauss-Newton iteration are e.g. Levenberg-Marquardt's method, or adapting the stepsize by using $\mu_p U^{-1}f$ instead of $U^{-1}f$ for some $\mu_p \in (0,1)$. One can ensure that the Gauss-Newton iteration step will increase the posterior density by modifying the stepsize.

Existence of a minimum $\tilde{\kappa}^*$ is guaranteed, if $\tilde{V}_\alpha = \mathbf{R}^{m-1}$, because the criterion $J_T(\kappa)$ is coercive. The chart $(\tilde{U}_\alpha, x_\alpha; \tilde{U}_\alpha \to \tilde{V}_\alpha)$ was such that the transition equation of the

general model was linear on it. Hence, the condition $\tilde{V}_\alpha = \mathbb{R}^{m-1}$ can only be satisfied if the metric is complete.

More precise conditions on existence of minima on given charts and on convergence of Gauss-Newton estimates remain to be established, as well as a quantitative evaluation of the advantages of Fahrmeir/Kaufmann's method in comparison to the ordinary Extended Kalman Filter and Smoother.

6.1.7 Observability for geodesic models

In the beginning of this section we introduced the random information at time t for a zero-curvature model, $\tilde{M}_t$, and the *conditional information* , defined by:

$$M_t(\kappa_t) = \boldsymbol{E}(\tilde{M}_t(\kappa_t)|\kappa_t, y_{-n}, \cdots, y_{t-1}), \tag{6.60}$$

which is equal to

$$M_t = 2Dr'(\rho_t)Dr(\rho_t) + Da'(\tilde{\kappa}_t)\frac{\phi_t\phi_t'}{e^{2r(\rho_t)}}Da(\tilde{\kappa}_t). \tag{6.61}$$

We are going to calculate an asymptotic expression for this conditional information. In order to do this, we have to combine corollary 4.14 and proposition 5.1. Let $(y_{t,T})_{0\leq t\leq RT}, T \in \mathbb{N}$ be a sequence of time-varying AR(n) processes with asymptotic relative running time ν and relative consideration time R, and satisfying assumptions 1,2,3,4,5 of theorem 4.13. According to this theorem, on the event that $\nu > R$ and conditional on the σ-algebra $\mathcal{B}_R$, the sequence of processes admits the Dahlhaus-spectral density

$$f(u, w|\theta) = \frac{\sigma(u)^2}{|1 - a_1(u)w - \cdots - a_n(u)w^n|^2}. \tag{6.62}$$

We define:

$$L_{n,u|\theta}(x) = \frac{1}{\sigma(u)}(x^n - a_1(u)x^{n-1} - \cdots - a_n(u)) \tag{6.63}$$

as the asymptotic instantaneous associated polynomials of the sequence of processes. Now, we can reformulate corollary 4.14 of chapter four as follows:

Corollary 6.5 On the event that $\nu > R$ and conditional on the σ-algebra $\mathcal{B}_R$, the autocovariances of the processes $(y_{t,T})_{0\leq t\leq RT}$ are asymptotically:

$$\begin{aligned} \lim_{T\uparrow\infty} \boldsymbol{Cov}^{\mathcal{B}_R}(y_{[uT-j],T}, y_{[uT-l],T}) &= \langle x^{j-1}, x^{l-1}\rangle_{L_{n,u|\theta},Fisher} \\ \text{for all } l,j,\ 1 \leq l,j \leq n; &\qquad \text{for all } u \in [0,R] \end{aligned} \tag{6.64}$$

Proof: Apply corollary 4.14 and equation (2.30) in chapter two to see that the left hand side is the shift invariant inner product, induced by $L_{n,u|\theta}$ on Π_{n-1}; apply proposition (5.1) to see that this is the same as the inner product on Π_{n-1}, induced by the asymptotic Fisher metric in the point $L_{n,u|\theta}$. □

On the chart $(\tilde{U}_\beta, x_\beta)$, introduced for the zero curvature model (6.18-6.20,6.7), we have $g_\beta^N = I_m$, hence the orthonormal vector fields F_i in the direction space $\mathcal{L}$ are the coordinate vector fields:

$$F_j = \frac{\partial}{\partial x_{\beta j}} = \frac{\partial \psi_\beta}{\partial \kappa^j}. \tag{6.65}$$

We represent the elements of U_n (and its tangent spaces) as polynomials, hence

$$\begin{aligned} L_n = \psi_\beta(\tilde{\kappa}, \rho) &= e^{-r(\rho)}(x^n - a_1(\tilde{\kappa})x^{n-1} - \cdots - a_n(\tilde{\kappa})) \\ F_j = \frac{\partial \psi_\beta}{\partial \tilde{\kappa}^j} &= e^{-r(\rho)}(-\frac{\partial a_1}{\tilde{\kappa}^j}(\tilde{\kappa})x^{n-1} - \cdots - \frac{\partial a_n}{\tilde{\kappa}^j}(\tilde{\kappa})) \\ F_m = \frac{\partial \psi_\beta}{\rho} &= \frac{\partial r}{\partial \rho}(\rho) L_n(x) \end{aligned} \tag{6.66}$$

According to proposition 5.1, $L_n \perp \Pi_{n-1}$ and $\langle L_n, L_n \rangle_{L_n} = 2$ if the asymptotic Fisher metric is used. Equations (6.61, 6.66) and corollary 6.5 yield immediately

Lemma 6.6 The expectation of the information matrix $M_{[uT]}$, conditional on the coefficient process asymptotically, becomes:

$$\lim_{T\uparrow\infty} \boldsymbol{E}^{\mathcal{B}_R} \tilde{M}_{[uT],T}(\kappa_{[uT],T}) = \lim_{T\uparrow\infty} \boldsymbol{E}^{\mathcal{B}_R} M_{[uT],T}(\kappa_{[uT],T}) = (\langle F_i, F_j \rangle_{\text{a.F.m.}})_{i,j} = H_\beta(\gamma(u)) \tag{6.67}$$

where H_β is the matrix on the coordinates ψ_β of the Riemannian metric induced on $N(q_0)$ by restricting the asymptotic Fisher metric to this manifold. □

Zero-curvature models exist for all metrics, including the asymptotic Fisher metric itself; for example, the geodesic models. If the geodesic tangent field is F, then, for a geodesic model based on the asymptotic Fisher metric we get the *asymptotic expected information at relative time point* $u \in [0, R]$:

$$\lim_{T\uparrow\infty} \boldsymbol{E}^{\mathcal{B}_R} \tilde{M}_{[uT],T}(\kappa_{[uT],T}) = \lim_{T\uparrow\infty} \boldsymbol{E}^{\mathcal{B}_R} M_{[uT],T}(\kappa_{[uT],T}) = \langle F, F \rangle_{\text{a.F.m.}} = 1. \tag{6.68}$$

Interpretation: *if we use the asymptotic Fisher metric in the Geodesic Model, then asymptotically (in the Dahlhaus-sense) we suppose that the contributions of the observations* $y_{[uT],T}$ $(u \in [0,1])$ *to the information about the changing parameters* κ *are equal.*□

Hence, the parametrization, induced by the asymptotic Fisher metric, stabilizes the expected contribution of each datapoint to the information about the coefficients. We have seen in chapter five that such a parametrization also stabilizes the variance of an ML estimator of the coefficient in the time-invariant AR(n) situation.

For other metrics the situation is different. For instance, in a model on AR_1 based on the Euclidean metric, this asymptotic expected information in $u \in [0, R]$ is $\frac{1}{1-a(u)^2}$, and if we use the complete metric introduced in chapter five (see equation 5.47), this is $1-a(u)^2$.

However, we have to note that asymptotically in the Dahlhaus sense, the data are more used to determine the hyperparameters than the coefficients.

The random informations $\tilde{M}_t$ and the conditional informations M_t are related to the concept of uniform complete observability (Jazwinski, 1970, chapter 7). This concept and the concept of uniform controllability play an important role in linear filtering theory. We illustrate them for *geodesic* models. Then, the linearized model (6.36) has two-dimensional states.

Let $a \in (0,1)$ and $\tau \in \mathbb{R}^+$. Consider the linear stochastic system with two-dimensional states and time-varying measurement matrices

$$\begin{aligned} x_{t+1} &= \mathcal{A}x_t + \mathcal{B}\lambda_t \\ z_t &= \begin{pmatrix} c_t & 0 \end{pmatrix} x_t + \sigma_t \epsilon_t, \quad \text{where} \\ \begin{pmatrix} \lambda_t \\ \epsilon_t \end{pmatrix} &\overset{i.i.d.}{=} \mathcal{N}(0, I_2) \text{ for all } t \in \mathbf{Z}_+; \\ & c_t \in \mathbb{R}; \quad \sigma_t \in \mathbb{R}^+; \end{aligned} \tag{6.69}$$

$$\mathcal{A} = \begin{pmatrix} 1 & a \\ 0 & a \end{pmatrix}; \quad \mathcal{B} = \begin{pmatrix} \tau \\ \tau \end{pmatrix}. \tag{6.70}$$

Let

$$M_t = \frac{c_t^2}{\sigma_t^2}.$$

Let $N \in \mathbb{N}$. For $t \in \mathbb{N}$ we define the *observability gramian over the times* $t+1, \cdots, t+N$, $\mathcal{J}(t+N, t+1)$, by:

$$\mathcal{J}(t+N,t+1) = \sum_{k=1}^{N} \mathcal{A}'^{k-N} \frac{\begin{pmatrix} c_{t+k} \\ 0 \end{pmatrix} \begin{pmatrix} c_{t+k} & 0 \end{pmatrix}}{\sigma_t^2} \mathcal{A}^{k-N} = \mathcal{A}'^{-N} \left(\sum_{k=1}^{N} \mathcal{A}'^{k} \begin{pmatrix} M_{t+k} & 0 \\ 0 & 0 \end{pmatrix} \mathcal{A}^k \right) \mathcal{A}^{-N}. \tag{6.71}$$

Jazwinski calls this matrix the *information matrix*. This term can be explained by the fact that, if $\tau = 0$, then there is an unbiased estimator of x_{t+N} from the data $y_{t+1}, \cdots, y_{t+N}$, for which the variance is minimal, and this minimal variance is equal to $\mathcal{J}(t+N, t+1)^{-1}$. Hence, $\mathcal{J}(t+N, t+1)$ coincides in this situation with the notion of information used, in the classical estimation theory (Jazwinski, 1970,chapter 7).
The linear system (6.70) is called *uniformly completely observable* (Jazwinski, 1970, chapter 7) if there exists $N \in \mathbb{N}$ and positive constants α, β such that

$$0 < \alpha I < \mathcal{J}(t+N, t+1) \leq \beta I, \text{ for all } t \in \mathbf{Z}_+. \tag{6.72}$$

For system (6.70) the *controllability gramian over the times* $t+1, \cdots, t+N$, is time-independent. It is defined as:

$$\mathcal{K}(t+N, t+1) = \begin{pmatrix} \mathcal{B} & \mathcal{A}\mathcal{B} & \cdots & \mathcal{A}^{N-1}\mathcal{B} \end{pmatrix} \begin{pmatrix} \mathcal{B}' \\ \mathcal{B}'\mathcal{A}' \\ \vdots \\ \mathcal{B}'\mathcal{A}'^{N-1} \end{pmatrix}. \tag{6.73}$$

The system is called *uniformly completely controllable* if there exists $N \in \mathbb{N}$ and positive constants α, β such that

$$0 < \alpha I < \mathcal{K}(t+N, t+1) \leq \beta I, \text{ for all } t \in \mathbf{Z}_+. \tag{6.74}$$

Clearly, system (6.70) is uniformly completely controllable: just take $N = 2$. In linear Kalman Filter theory, these two concepts are used to prove asymptotic stability of the filter, and to bound the error covariances. We shall use these concepts too, although the geodesic model is *non-linear* in general. The concepts will be helpful in bounding the matrix U of the Gauss-Newton iteration in the point $\hat{\tilde{\kappa}}^p$ away from zero.

Now we shall investigate a condition for uniform complete observability.

Lemma 6.7 Suppose that for some $N \in \mathbb{N}$ there are positive numbers μ and $\tilde{\mu}$ such that

$$0 < \mu \leq \sum_{k=0}^{N} M_{t+k} \leq \tilde{\mu} \text{ for all } t \in \mathbb{N} \tag{6.75}$$

Then, the system (6.70) is uniformly completely observable.

Proof: First note that

$$\begin{aligned} \mathcal{A}^k &= \begin{pmatrix} 1 & a \\ 0 & a \end{pmatrix}^k = \begin{pmatrix} 1 & b_k \\ 0 & a^k \end{pmatrix} \text{ where} \\ b_k &= \sum_{i=1}^{k} a^i \text{ if } k > 0; \\ b_k &= -\sum_{i=0}^{k-1} (\frac{1}{a})^i \text{ if } k < 0 \text{ and } a > 0. \end{aligned} \tag{6.76}$$

The $(b_k)_{k \in \mathbb{N}}$ are *strictly* increasing. From equation(6.76) it follows that

$$\mathcal{A}'^{2N} \mathcal{J}(t+2N, t) \mathcal{A}^{2N} = \sum_{k=1}^{2N} \mathcal{A}'^k \begin{pmatrix} M_{t+k} & 0 \\ 0 & 0 \end{pmatrix} \mathcal{A}^k = \begin{pmatrix} \sum_{k=1}^{2N} M_{t+k} & \sum_{k=1}^{2N} M_{t+k} b_k \\ \sum_{k=1}^{2N} b_k M_{t+k} & \sum_{k=1}^{2N} M_{t+k} b_k^2 \end{pmatrix} =$$

$$= \sum_{k=1}^{2N} M_{t+k} \begin{pmatrix} 1 & 0 \\ \sum_{k=1}^{2N} p_k b_k & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \sum_{k=1}^{2N} p_k b_k^2 - (\sum_{k=1}^{2N} p_k b_k)^2 \end{pmatrix} \begin{pmatrix} 1 & \sum_{k=1}^{2N} p_k b_k \\ 0 & 1 \end{pmatrix}. \tag{6.77}$$

In this last equation,

$$p_k = \frac{M_{t+k}}{\sum_{j=0}^{2N} M_{t+j}}, \text{ hence } \sum_{k=0}^{2N} p_k = 1 \text{ and } p_k \geq 0.$$

Accordingly, $(p_k)_{0 \leq k \leq 2N}$ is a probability measure on the set $\{0, 1, \cdots, 2N\}$, and $V = \sum_{k=1}^{2N} p_k b_k^2 - (\sum_{k=1}^{2N} p_k b_k)^2$ is the variance of the variable $(b_k)_{0 \leq k \leq 2N}$. This variance is only zero if $b_k = b_j$ a.s., and, as $(b_k)_{0 \leq k \leq 2N}$ is strictly increasing, this is only possible if one of the $p_k = 1$ and the others are zero. Now, $\mu \leq \sum_{k=1}^{N} M_{t+k} \leq \tilde{\mu}$ for all $t \in \mathbb{N}$, hence there is a $k_1, 1 \leq k_1 \leq N$, such that $M_{t+k_1} \geq \frac{\mu}{N}$. Also $\mu \leq \sum_{k=N+1}^{2N} M_{t+k} \leq \tilde{\mu}$, hence there is

a number $k_2, N+1 \leq k_2 \leq 2N$, such that $M_{t+k_2} \geq \frac{\mu}{N}$. Consequently, the numbers p_{k_1} and p_{k_2} satisfy

$$p_{k_1} \geq q \text{ and } p_{k_2} \geq q, \text{ where } q = \frac{\mu}{2N\tilde{\mu}}. \tag{6.78}$$

If we minimize V over the set

$$\{(p_0, \cdots, p_{2N}) | \sum_{k=0}^{2N} p_k = 1 \wedge p_k \geq 0 \; \forall k \wedge \exists_{k_1,k_2}[0 < k_1 < k_2 \leq 2N \wedge p_{k_1} \geq q \wedge p_{k_2} \geq q]\},$$

then we get a positive minimum, because the set is compact, V is a continuous function of $p_0, \cdots, p_{2N}$, and V is strictly positive on this set, as we argued above. The minimum δ is only a function of $\frac{\mu}{\tilde{\mu}}, N$ and the number a. Hence we have:

$$0 < \delta < V < \Delta. \tag{6.79}$$

Next, we derive inequalities for the eigenvalues ρ of

$$G = \begin{pmatrix} 1 & 0 \\ \sum_{k=1}^{2N} p_k b_k & 1 \end{pmatrix} \begin{pmatrix} 1 & \sum_{k=1}^{2N} p_k b_k \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & \sum_{k=1}^{2N} p_k b_k \\ \sum_{k=1}^{2N} p_k b_k & (\sum_{k=1}^{2N} p_k b_k)^2 + 1 \end{pmatrix}.$$

It is immediate that $(\sum_{k=1}^{2N} p_k b_k)^2 = \rho + \frac{1}{\rho} - 2$, hence

$$\frac{1}{b_{2N}^2 + 2} \leq \rho \leq b_{2N}^2 + 2.$$

This concludes the proof. □

Note that the linearized model(6.36) for a geodesic model has the form of system (6.70). If model (6.36) is uniformly completely observable, where $\mathcal{J}(t+N, t+1)$ satisfies inequality (6.72) and the controlability gramian $\mathcal{K}(t+N, t+1)$ satisfies inequality (6.74), then Lemma 7.1. in Jazwinski (1970) indicates that the filter error covariances $\Sigma_{t|t}$ of the states x_t are bounded from above by:

$$\Sigma_{t|t} \leq (\frac{1+\alpha\beta}{\alpha})I$$

for all $t \geq N$. As $\Sigma_{1|1} > 0$ we can also apply the subsequent Lemma 7.2 of Jazwinski: for all $t \geq N+1$ these covariances are bounded from below by:

$$\Sigma_{t|t} \geq \frac{\alpha}{1+\alpha\beta} I.$$

Moreover, one can write the transition equation for the filter estimates $x_{t|t}$ of the states of model (6.36) as follows:

$$x_{t|t} = \Sigma_{t|t}\Sigma_{t|t-1}^{-1} \mathcal{A} x_{t-1|t-1} + \Sigma_{t|t} \frac{C_t'}{\sigma_t^2} z_t \tag{6.80}$$

Denote the transition matrix by G_t:

$$G_t = \Sigma_{t|t}\Sigma_{t|t-1}^{-1} \mathcal{A} \tag{6.81}$$

In case of uniform complete observability and uniform complete controllability, the system

$$x_t = G_t x_{t-1} \tag{6.82}$$

is uniformly asymptotically stable, according to Theorem 7.4, Jazwinski (1970), i.e.

$$||G_t G_{t-1} \cdots G_{k+1} G_k|| \leq \nu e^{-\tilde{\nu}(t-k)} \quad \text{for all } t \geq k \tag{6.83}$$

where ν and $\tilde{\nu}$ are positive constants, of course (only) depending on α, β, N and a. Now we can reorder the product of these transition matrices:

$$G_t G_{t-1} \cdots G_k = \Sigma_{t|t} F'_{t-1} \cdots F'_{k-1} \Sigma^{-1}_{k-1|k-1},$$

where the F_t are the smoothing gain operators, defined in (6.42). Consequently, one has, for all $t \geq k \geq N$:

$$||F_k F_{k+1} \cdots F_t|| \leq (\frac{1+\alpha\beta}{\alpha})^2 \nu e^{-\tilde{\nu}(t-k)},$$

Look at the total variance S of the states $x_0, \cdots, x_T$, conditional on $z_0, \cdots, z_T$. As $S_{t\ t+k} = F_t \cdots F_{t+k-1} \Sigma_{t+k|T}$ and $\Sigma_{t+k|T} \leq \Sigma_{t+k|t+k} \leq \frac{1+\alpha\beta}{\alpha}$ the variance S is bounded from above by:

$$P = \frac{2\nu}{1-e^{-\tilde{\nu}}} (\frac{1+\alpha\beta}{\alpha})^3 I.$$

Consequently, also the variance $\mathcal{V}$ of the κ_t, given $z_0, \cdots, z_T$, satisfies $\mathcal{V} \leq P$.
This gives a lower bound for the gain matrix U:

$$U \geq P^{-1} \tag{6.84}$$

and this bound is only a function of α, β, N and a (and not of T).

The concepts of observability and controllability not only enable us to say something about the smallest eigenvalue of the gain matrix. We deduce also a statement about the errors in the initial values.
If model (6.36) is uniformly completely observable and controllable, an error in the estimate $\Sigma_{0|-1}$ of the error variance of the initial coefficient coordinate κ_0 will gradually be forgotten, i.e. will go exponentially to zero (theorem 7.5, Jazwinski (1970)).

6.1.8 Estimating the noise levels

We treat the second half "Gauss"-Newton iteration step and an alternative to it in this subsection. In this half step, one considers the coefficients $a_i(\tilde{\kappa})$ as being fixed, and the noise levels are updated. The log noise levels $r(\kappa_t)$ are multiples of one of the coordinates of the parametrization $r(\kappa_t) = \mu\rho_t$. Consequently, $Dr(\kappa_t) = \begin{pmatrix} 0 & 0 & \cdots & 0 & \mu \end{pmatrix}$. The score functions m_t with respect to the log noise level are:

$$m_t = 1 - e^{-2\mu\rho_t}(y_t - \phi'_t a(\tilde{\kappa}_t))^2. \tag{6.85}$$

The terms c_t remain as given in (6.25), but are one-dimensional now. The (one-dimensional) block of the conditional information M_t, with respect to the log noise level,

is just $2\mu^2$. Consequently, the block of the gain matrix U, to be used in this second half iteration step, is almost a Toeplitz matrix. The important difference with a Toeplitz matrix $\tilde{U}$ is the block of size 2×2 at the bottom at the right hand corner. The unimportant difference is the element at the top at the left hand corner. U is the same matrix as one gets in twice, deriving the log posterior density of $\rho_0, \cdots, \rho_T$ of the model

$$\begin{aligned} x_t &= \begin{pmatrix} \rho_t \\ \theta_t \end{pmatrix} \\ x_{t+1} &= \begin{pmatrix} 1 & a \\ 0 & a \end{pmatrix} x_t + \begin{pmatrix} \frac{1}{p} \\ \frac{1}{p} \end{pmatrix} \lambda_{t-1} \\ z_t &= \begin{pmatrix} 2\mu & 0 \end{pmatrix} x_t + \eta_t \\ x_0 &\stackrel{d}{=} \mathcal{N}(\begin{pmatrix} \rho_{-1} \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{Q}{p^2} & \frac{Q}{p^2} \\ \frac{Q}{p^2} & \frac{Q}{p^2} \end{pmatrix}) \\ (\lambda_t)_{t\in\mathbf{Z}} \quad ; \quad & (\eta_t)_{t\in\mathbf{Z}} \text{ independent sequences of} \\ & \text{Gaussian unit variance white noise.} \end{aligned} \tag{6.86}$$

(Note that we have used lower case letters a and p, instead of the capitals A and P in equations (6.28)).
This is a time-invariant, detectable and controllable system for all values of a. Consequently, if T tends to infinity,[2] then for all $k > 1$ the covariances $\Sigma_{T-k|T-k}$ tend to the only self-adjoint non-negative definite solution Σ_∞ of the Riccati equation

$$(I + (\mathcal{A}\Sigma_\infty\mathcal{A}' + \mathcal{B}\mathcal{B}')\frac{\mathcal{C}'\mathcal{C}}{\sigma^2})\Sigma_\infty = \mathcal{A}\Sigma_\infty\mathcal{A}' + \mathcal{B}\mathcal{B}' \tag{6.87}$$

where

$$\mathcal{A} = \begin{pmatrix} 1 & a \\ 0 & a \end{pmatrix}; \quad \mathcal{B} = \begin{pmatrix} \frac{1}{p} \\ \frac{1}{p} \end{pmatrix}; \quad \mathcal{C} = \begin{pmatrix} 2\mu & 0 \end{pmatrix}; \quad \sigma = 1,$$

and the filter gain matrices G_{T-k} tend to the asymptotically stable matrix G_∞, where

$$G_\infty = (I - \Sigma_\infty \frac{\mathcal{C}'\mathcal{C}}{\sigma^2})\mathcal{A}.$$

Statement. $\Sigma_\infty > 0$.
Proof: From the last equation of (6.47), it is clear that $\Delta_T \leq U_{TT} = 2\mu^2 + p^2$ for all T, and also $\Delta_{T-1} \leq U_{T-1\,T-1} = 2\mu^2 + p^2(2+2a+a^2)$, hence the sequences $(\Delta_T^{-1})_{T\in\mathbf{N}}$, $(\Delta_{T-1}^{-1})_{T\in\mathbf{N}}$ are bounded away from zero. The filter error covariances $\Sigma_{T|T} = D_T$ are also bounded away from zero, according to equation (6.51). □

Consequently, also the smoothing gain matrices F_{T-k} tend to an asymptotically stable matrix F_∞, where

$$F_\infty = \Sigma_\infty\mathcal{A}'(I - \frac{\mathcal{C}'\mathcal{C}}{\sigma^2}\Sigma_\infty)\Sigma_\infty^{-1}.$$

[2]Here we are not performing Dahlhaus asymptotics. The variables a and p are assumed to be independent of T.

Hence, the smoothing gain matrices $(B_{T-k-1})_{T\in\mathbb{N}}$, and $(C_{T-k-1})_{T\in\mathbb{N}}$ ($k \in \mathbb{N}$) have a limit B_∞ and C_∞ respectively, where

$$x^2 - B_\infty x + C_\infty x \text{ is stable,} \tag{6.88}$$

because of equation (6.49). Equations (6.47) show that all sequences $(B_{T-k})_{T\in\mathbb{N}}$, $(B_{T-1})_{T\in\mathbb{N}}$, $(C_{T-k})_{T\in\mathbb{N}}$, $(\Delta_{T-k})_{T\in\mathbb{N}}$, $(\Delta_{T-1})_{T\in\mathbb{N}}$, $(\Delta_T)_{T\in\mathbb{N}}$ converge. One could work with these limits, e.g. for larger values of T, instead of with the genuine $B_{T-k}, B_{T-1}, C_{T-k}, \Delta_{T-k}, \Delta_{T-1}, \Delta_T$ without harming the results. This is in the same line as working with the time-invariant Kalman filter for the linear model (6.86), which is asymptotically optimal. The limit values can be computed from equations (6.47).
We have also a result about the smallest eigenvalue $\lambda_{\min}(U)$ of U.

Statement. $\lim_{T\uparrow\infty} \lambda_{\min}(U) = 2\mu^2$

Proof: Note that $\tilde{U}$ is part of the covariance matrix of an MA(2)-process which is stationary and regular, and, hence should have spectral density $g(\phi) = \Delta_\infty|(e^{i2\phi} - B_\infty e^{i\phi} + C_\infty)|^2$. In the limit, the smallest eigenvalue of $\tilde{U}$ will be the minimum of g. From equations (6.47), we derive that

$$\Delta_\infty(1 - B_\infty + C_\infty)^2 = 2\mu^2.$$

The left hand side is just $g(0)$. Consequently, the smallest eigenvalue of $\tilde{U}$ is smaller or equal to $2\mu^2$ in the limit. On the other hand, from equations (6.28) and (6.29) we have:

$$2\mu^2 I \leq U \leq \tilde{U}. \qquad \square \tag{6.89}$$

So, Fahrmeir/Kaufmann's method and subsequent iterations, as described for the first half step, can be carried out. As we don't have a (non-linear) least squares problem in the strict sense of the word, we cannot refer to the theorem of Dennis and Schnabel to analyze the "Gauss-Newton"-method. We have to appeal on a theorem of Kantorovich and Akilov (XVIII, 2.1; 2.2), in which conditions are given for local existence, uniqueness of a zero of f, and convergence of a Newton sequence with modified gain matrix to that zero. However, because of the large non-linearities involved, one can hardly expect that these conditions will be satisfied. But again, if we are close enough to a minimum κ^* of the criterion, and $Df(\kappa^*)$ is positive definite, the strict Newton method (with $U = Df(\hat{\kappa}^p)$) yields a convergent sequence. It can be carried out by means of the Fahrmeir/Kaufmann's procedure plus iterations. If we do so, we should not separate the determination of noise levels and coefficients.

There is an alternative to the G-N second half step. It consists in approximating the log-likelihood terms l_t in the criterion $J_T(\rho)$. (The coefficients $a(\tilde{\kappa})$ are again considered to be known). The idea to approximate $l_t = \mu\rho_t + \frac{1}{2}e^{-2\mu\rho_t}(y_t - \phi_t' a(\tilde{\kappa}_t))^2$ can be found in Kitagawa and Gersch (1985).
First a remark. If $(\epsilon_t)_{t\in\mathbb{Z}}$ is Gaussian white noise with unit variance, then also $(\tilde{\eta}_t)_{t\in\mathbb{Z}}$, defined by:

$$\tilde{\eta}_t = \log(\frac{\epsilon_t^2}{2})$$

are mutually independent identically distributed variables, but their distribution is not Gaussian anymore. It is not difficult to calculate the moment generating function of $\tilde{\eta}_t$:

$$\boldsymbol{E}\exp(z\tilde{\eta}_t) = \frac{\Gamma(z+\frac{1}{2})}{\Gamma(\frac{1}{2})}, \tag{6.90}$$

where Γ is the Gamma function. Let γ be Euler's number. It follows from (6.90) that

$$\begin{aligned} \boldsymbol{E}\tilde{\eta}_t &= -\gamma - \log(4) \\ \mathrm{Var}(\tilde{\eta}_t) &= \frac{5\pi^2}{9} := R. \end{aligned} \tag{6.91}$$

If we define

$$\eta_t := \log(2\epsilon_t^2) + \gamma = \tilde{\eta}_t + \gamma + \log(4)$$

then, it is clear that the sequence $(\eta_t)_{t\in\mathbf{Z}}$ is zero mean white noise with variance equal to $R = \frac{5\pi^2}{9}$.
Conditional on coefficient coordinates $\tilde{\kappa}_s$ and noise level coordinates ρ_s, the sequence

$$\epsilon_t = e^{-\mu\rho_t}(y_t - \phi_t' a(\tilde{\kappa}_t))$$

is Gaussian unit variance white noise. Set $z_t := \log 2(y_t - a(\tilde{\kappa}_t))^2 + \gamma$. Then, conditional on $\tilde{\kappa}_s$, ρ_s,

$$z_t = 2\mu\rho_t + \eta_t,$$

(η_t as defined in the remark.) As we want to minimize the criterion J_T only with respect to ρ_t, we can as well minimize the criterion $\tilde{J}_T$, where the log-likelihood terms l_t have been replaced by:

$$\tilde{l}_t = l_t - \frac{1}{2}z_t.$$

A comparison of the conditional distribution, given the $\tilde{\kappa}_s$, ρ_s for $\tilde{l}_t$ with the conditional distribution for $\frac{(z_t-2\mu\rho_t)^2}{R}$, makes it reasonable to replace the log-likelihood terms l_t in the criterion J_T altogether by $\hat{l}_t = \frac{(z_t-2\mu\rho_t)^2}{R}$. We obtain the new criterion $\mathcal{J}_T$. It would be nice if minimization of $\mathcal{J}_T$ is asymptotically the same as minimization of the old criterion J_T or $\tilde{J}_T$. However, we doubt that this is true, because for given $r^T = (r_0^T, \cdots, r_T^T)$ the expectation (not conditional on r^T)

$$\frac{1}{T+1}\boldsymbol{E}(\tilde{J}_T(r^T) - \mathcal{J}_T(r^T))$$

does not tend to zero.
The new criterion is just the $-\log$ joint density of the linear model (6.86) (with the only exception that the measurement error covariance is R instead of 1). For minimization of this criterion, not the whole Fahrmeir/Kaufmann procedure is necessary, as we explained in the last subsection; only Kalman Filtering and Smoothing once is sufficient. Note that model (6.86) relates the a priori non-stationary smoothly integrated randomwalk ρ_t to an asymptotically stationary AR(2) process a posteriori with a spectral density $\frac{\Delta_\infty^{-1}}{|e^{2i\phi} - B_\infty e^{i\phi} + C_\infty|^2}$.

6.2 Approximating the parametrization

We have proposed an identification method for zero-curvature models. We implicitly assumed that the parametrization, on which the transition equation is linear, is available in an analytical form, for example, the formulae of some pencils of geodesics. We have given some examples of pencils of geodesics for the asymptotic Fisher metric in chapter five. However, the purpose of this section is to show that it is not necessary for identification to have such an analytic expression of the parametrization at one's disposal. It is only necessary to have an explicit formula for the metric of the k-dimensional manifold M on *some* set of charts.

According to corollary 3.11, the direction space $\mathcal{L}$ of a zero-curvature model satisfies

$$\nabla_X Y = 0 \text{ for all } X, Y \in \mathcal{L}, \tag{6.92}$$

hence all vector fields in $\mathcal{L}$ are tangent fields of geodesics and the distribution $\mathcal{D}$, generated by $\mathcal{L}$, is totally geodesic and has zero curvature.

If one does not have an analytic expression of this chart, on which the transition equation is linear, then approximation is necessary. In order not to complicate the situation more than needed, we suppose that the manifold M is covered by one chart $(\tilde{U}_\beta = M, x_\beta)$ that is adapted to the distribution $\mathcal{D}$, and on which the transition equation is completely linear. We also suppose that M is covered by one *other* chart $(\tilde{U}_\alpha = M, x_\alpha)$, on which an explicit form of the metric is known. In case M is an open part of U_n, the latter chart can be the natural one:

$$M \ni q = e^{-r}(x^n - a_1 x^{n-1} + \cdots - a_n) \mapsto (a_1, \cdots, a_n, r) \in AR_n \times \mathbb{R} \subset \mathbb{R}^{n+1}.$$

The first chart $(M, x_\beta : M \to V_\beta \subset \mathbb{R}^k)$ can be expressed in terms of the latter chart $(M, x_\alpha : M \to V_\alpha \subset \mathbb{R}^k)$ by means of the coordinate transformation $\psi = x_\alpha \circ x_\beta^{-1} : V_\beta \to V_\alpha$, which is a diffeomorphism. One can view ψ as an (indirect) parametrization of M. The coordinates of the vector $\frac{\partial \psi}{\partial \kappa_i}$ are the coefficients of the tangent vector $\frac{\partial}{\partial x_{\beta\, i}}$ with respect to the moving coordinate basis $(\frac{\partial}{\partial x_{\alpha\, j}})_{1 \le j \le k}$. Accordingly, we can write $\frac{\partial \psi}{\partial \kappa_i}$ instead of $\frac{\partial}{\partial x_{\beta\, i}}$. Hence, our problem is to approximate the function ψ and its derivatives.

The chart (M, x_β) is adapted to the m-dimensional distribution $\mathcal{D}$. This distribution has integral manifolds $N(q)$ through every point $q \in M$. The adaptedness of the chart means that the last $k - m$ components $\kappa_{m+1}, \cdots, \kappa_k$ of x_β are constant on any integral manifold $N(q)$, and that the first m components $\kappa_1, \cdots \kappa_m$ can be used as coordinates on that manifold, i.e.

$$N(q) = \{p \in M | (x_\beta(p))_{m+1} = (x_\beta(q))_{m+1}, \cdots, (x_\beta(p))_k = (x_\beta(q))_k\}.$$

In the sequel, we will denote these last $k - m$ components by $\zeta_1, \cdots, \zeta_{k-m}$ instead of $\kappa_{m+1}, \cdots, \kappa_k$, and use κ exclusively to denote the first m coordinates of x_β. Accordingly, we will write:

$$Z_i = \frac{\partial \psi}{\partial \zeta_i}; \tag{6.93}$$

$$F_j = \frac{\partial \psi}{\partial \kappa_j}. \tag{6.94}$$

The vector fields F_j $(1 \leq j \leq m)$ are elements of $\mathcal{L}$. The transition equation is completely linear on the chart x_β, and we can choose the chart, such that for any $q \in M$ the m-dimensional matrix $g^N_\beta = I$, constant on the manifold $N(q)$. (This is the same as choosing $F_1, \cdots, F_m$ to be an orthonormal basis of eigen vectors of the symmetric tensor V in $\mathcal{L}$). Consequently, the inner products

$$\langle \frac{\partial \psi}{\partial \kappa_i}, \frac{\partial \psi}{\partial \kappa_j} \rangle = \langle F_i, F_j \rangle = \delta_{ij},$$

are constant on $N(q)$. As the vector fields F_i are elements of $\mathcal{L}$, they satisfy:

$$\nabla_{F_i} F_j = 0. \tag{6.95}$$

The vector fields Z_l, F_j are coordinate vector fields, hence they satisfy $[Z_l, F_j] = 0$. This, together with equation (6.95), imply the following relation:

$$\nabla_{F_i} \nabla_{F_j} Z_l = R(F_i, Z_l) F_j \tag{6.96}$$

where $R(X,Y)Z$ is the curvature tensor [3]). We shall now develop an approximation method for ψ, and its derivatives based on recursions using

$$F_i, \quad Z_l \quad \text{and } W_{jl} := \nabla_{F_j} Z_l.$$

To that end, we need analytic expressions for the *Christoffel* and *curvature* symbols Γ^s_{ij} and R^s_{rij} of the metric with respect to the chart $(\tilde{U}_\alpha, x_\alpha)$. These symbols are implicitly defined as follows:
if $U = \sum_{i=1}^k u_i \frac{\partial}{\partial x_{\alpha\, i}}$; $V = \sum_{i=1}^k v_i \frac{\partial}{\partial x_{\alpha\, i}}$ and $W = \sum_{i=1}^k w_i \frac{\partial}{\partial x_{\alpha\, i}}$ then

$$\nabla_U V = \sum_{s=1}^k (Uv_s + \sum_{i,j=1}^k \Gamma^s_{ij} u_i v_j) \frac{\partial}{\partial x_{\alpha\, s}}; \tag{6.97}$$

$$R(U,V)W = \sum_{s=1}^k (\sum_{i,j,r=1}^k R^s_{rij} u_i v_j w_r) \frac{\partial}{\partial x_{\alpha\, s}}. \tag{6.98}$$

If $H := H_\alpha$ is the matrix of the metric on the chart $(\tilde{U}_\alpha, x_\alpha)$, then we have the following relations between this metric and the Christoffel and curvature symbols:

$$\Gamma^s_{ij} = \frac{1}{2} \sum_{l=1}^k (H^{-1})_{sl} (-\frac{\partial H_{ij}}{\partial x_{\alpha\, l}} + \frac{\partial H_{jl}}{\partial x_{\alpha\, i}} + \frac{\partial H_{li}}{\partial x_{\alpha\, j}}) \tag{6.99}$$

$$R^m_{rij} = \frac{\partial \Gamma^m_{jr}}{\partial x_{\alpha\, i}} - \frac{\partial \Gamma^m_{ir}}{\partial x_{\alpha\, j}} + \sum_{l=1}^k (\Gamma^l_{jr} \Gamma^m_{il} - \Gamma^l_{ir} \Gamma^m_{jl}). \tag{6.100}$$

The idea is that, if we made an initial good guess of $\zeta_{0,i} = \zeta_{1,i} = \zeta_{2,i} = \cdots$, the differences $\hat{\zeta}_{t+1,i} - \hat{\zeta}_{t,i}$ will be small in comparison with the differences $\hat{\kappa}_{t+1} - \hat{\kappa}_t$ [4].). So, in a Taylor

[3] A vector field Z along a geodesic with tangent field F, such that $\nabla_F \nabla_F Z = R(F,Z)F$, is called *Jacobian.*

[4] In the Fahrmeir/Kaufmann method, we have considered $\zeta_{0,i}$ $(1 \leq i \leq k-m)$ as hyperparameters, hence then the differences $\hat{\zeta}_{t+1,i} - \hat{\zeta}_{t,i}$ will be automatically zero and the approximation algorithm presented below will be much simpler. Still we need the derivatives with respect to ζ also in that case if we want to determine these hyperparameters $\zeta_{0,1}, \cdots, \zeta_{0,k-m}$ by maximizing the likelihood or posterior density by means of a Gauss-Newton iteration, see next section 6.3

series expansion of order 2 of $\psi(\hat{\kappa}_{t+1}, \hat{\zeta}_{t+1})$ around $(\hat{\kappa}_t, \hat{\zeta}_t)$, quadratic terms in the first type of differences will be neglected. In the Taylor series expansions of order 1 of the vector fields $Z_l(\hat{\kappa}_{t+1}, \hat{\zeta}_{t+1})$ and $W_{jl}(\hat{\kappa}_{t+1}, \hat{\zeta}_{t+1})$ around $(\hat{\kappa}_t, \hat{\zeta}_t)$, these differences will be neglected altogether. Then we obtain:

$$\begin{aligned}
\text{if } F_i &= \sum_{s=1}^{k} F_{i|s} \frac{\partial}{\partial x_{\alpha\, s}}; \quad Z_l = \sum_{s=1}^{k} Z_{l|s} \frac{\partial}{\partial x_{\alpha\, s}} \text{ and } W_{jl} = \sum_{s=1}^{k} W_{jl|s} \frac{\partial}{\partial x_{\alpha\, s}} \text{ then} \\
0 = \nabla_{F_i} F_j \;\Rightarrow\; \frac{\partial^2 \psi}{\partial \kappa_i \partial \kappa_j} &= - \begin{pmatrix} \sum_{s,r=1}^{k} \Gamma^1_{sr} F_{i|s} F_{j|r} \\ \vdots \\ \sum_{s,r=1}^{k} \Gamma^k_{sr} F_{i|s} F_{j|r} \end{pmatrix} \qquad (6.101) \\
W_{jl} &:= \nabla_{F_j} Z_l \Rightarrow \\
\frac{\partial^2 \psi}{\partial \kappa_j \partial \zeta_l} &= \begin{pmatrix} W_{jl|1} \\ \vdots \\ W_{jl|k} \end{pmatrix} - \begin{pmatrix} \sum_{s,r=1}^{k} \Gamma^1_{sr} F_{j|s} Z_{l|r} \\ \vdots \\ \sum_{s,r=1}^{k} \Gamma^k_{sr} F_{j|s} Z_{l|r} \end{pmatrix}; \qquad (6.102) \\
\nabla_{F_i} W_{jl} &= R(F_i, Z_l) F_j \Rightarrow \\
\frac{\partial W_{jl}}{\partial \kappa_i} &= \begin{pmatrix} \sum_{s,r,p=1}^{k} R^1_{srp} F_{i|s} Z_{l|r} F_{j|p} \\ \vdots \\ \sum_{s,r,p=1}^{k} R^k_{srp} F_{i|s} Z_{l|r} F_{j|p} \end{pmatrix} - \begin{pmatrix} \sum_{s,r=1}^{k} \Gamma^1_{sr} F_{i|s} W_{jl|r} \\ \vdots \\ \sum_{s,r=1}^{k} \Gamma^k_{sr} F_{i|s} W_{jl|r} \end{pmatrix}. \qquad (6.103)
\end{aligned}$$

The Taylor series approximations now become:

$$\begin{aligned}
\psi(\hat{\kappa}_{t+1}, \hat{\zeta}_{t+1}) &= \psi(\hat{\kappa}_t, \hat{\zeta}_t) + \sum_{i=1}^{m} F_i(\hat{\kappa}_t, \hat{\zeta}_t)(\hat{\kappa}_{t+1,i} - \hat{\kappa}_{t,i}) + \\
&\quad + \sum_{l=1}^{k-r} Z_l(\hat{\kappa}_t, \hat{\zeta}_t)(\hat{\zeta}_{t+1,l} - \hat{\zeta}_{t,l}) + \\
&\quad + \sum_{i=1}^{m} \sum_{l=1}^{k-r} \frac{\partial^2 \psi}{\partial \kappa_i \partial \zeta_l}(\hat{\kappa}_t, \hat{\zeta}_t)(\hat{\kappa}_{t+1,i} - \hat{\kappa}_{t,i})(\hat{\zeta}_{t+1,l} - \hat{\zeta}_{t,l}) + \\
&\quad + \frac{1}{2} \sum_{i,j=1}^{m} \frac{\partial^2 \psi}{\partial \kappa_i \partial \kappa_j}(\hat{\kappa}_t, \hat{\zeta}_t)(\hat{\kappa}_{t+1,i} - \hat{\kappa}_{t,i})(\hat{\kappa}_{t+1,j} - \hat{\kappa}_{t,j}) + \\
&\quad + O(|| \begin{pmatrix} \kappa \\ \zeta \end{pmatrix} ||^3) + O(||\zeta||^2) \qquad (6.104) \\
F_i(\hat{\kappa}_{t+1}, \hat{\zeta}_{t+1}) &= F_i(\hat{\kappa}_t, \hat{\zeta}_t) + \sum_{j=1}^{m} \frac{\partial^2 \psi}{\partial \kappa_i \partial \kappa_j}(\hat{\kappa}_t, \hat{\zeta}_t)(\hat{\kappa}_{t+1,j} - \hat{\kappa}_{t,j}) + \\
&\quad + \sum_{l=1}^{k-r} \frac{\partial^2 \psi}{\partial \kappa_i \partial \zeta_l}(\hat{\kappa}_t, \hat{\zeta}_t)(\hat{\zeta}_{t+1,l} - \hat{\zeta}_{t,l}) + O(|| \begin{pmatrix} \kappa \\ \zeta \end{pmatrix} ||^2) \qquad (6.105) \\
Z_l(\hat{\kappa}_{t+1}, \hat{\zeta}_{t+1}) &= Z_l(\hat{\kappa}_t, \hat{\zeta}_t) + \sum_{i=1}^{m} \frac{\partial^2 \psi}{\partial \kappa_i \partial \zeta_l}(\hat{\kappa}_t, \hat{\zeta}_t)(\hat{\kappa}_{t+1,i} - \hat{\kappa}_{t,i}) + \\
&\quad + O(|| \begin{pmatrix} \kappa \\ \zeta \end{pmatrix} ||^2) + O(||\zeta||) \qquad (6.106)
\end{aligned}$$

$$W_{jl}(\hat{\kappa}_{t+1},\hat{\zeta}_{t+1}) = W_{jl}(\hat{\kappa}_t,\hat{\zeta}_t)+\sum_{i=1}^{m}\frac{\partial W_{jl}}{\partial \kappa_i}(\hat{\kappa}_t,\hat{\zeta}_t)(\hat{\kappa}_{t+1,i}-\hat{\kappa}_{t,i})+$$
$$+O(||\begin{pmatrix}\kappa\\ \zeta\end{pmatrix}||^2)+O(||\zeta||). \qquad (6.107)$$

Illustration. For the a.F.m. on AR_2, the matrix of the metric H, and the *Christoffel-* and *curvature-* symbols Γ^s_{ij} and R^s_{rij} with respect to the natural coordinates

$$q = x^2 - a_1x - a_2 \mapsto (a_1,a_2)$$

were computed with the aid of "Mathematica", a computer algebra program. We show the results below:

We use the following notation:

$$\begin{aligned}
d &= (1+a_2)(1-a_1-a_2)(1+a_1-a_2);\\
\Gamma^s &= \begin{pmatrix}\Gamma^s_{11} & \Gamma^s_{12}\\ \Gamma^s_{21} & \Gamma^s_{22}\end{pmatrix} \text{ and}\\
R^m_s &= \begin{pmatrix}R^m_{s11} & R^m_{s12}\\ R^m_{s21} & R^m_{s22}\end{pmatrix}\\
&\text{with this notation,}\\
H &= \frac{1}{d}\begin{pmatrix}1-a_2 & a_1\\ a_1 & 1-a_2\end{pmatrix}\\
\Gamma^1 &= \frac{a_2}{d}\begin{pmatrix}a_1 & 1-a_2\\ 1-a_2 & a_1\end{pmatrix}\\
\Gamma^2 &= \frac{1}{d}\begin{pmatrix}1-a_2 & a_1\\ a_1 & a_1^2+a_2-a_2^2\end{pmatrix}\\
R^1_1 &= \frac{1}{d}\begin{pmatrix}0 & -\frac{a_1}{1+a_2}\\ \frac{a_1}{1+a_2} & 0\end{pmatrix};\quad R^1_2=\frac{1}{d}\begin{pmatrix}0 & -\frac{1-a_2}{1+a_2}\\ \frac{1-a_2}{1+a_2} & 0\end{pmatrix}\\
R^2_1 &= \frac{1}{d}\begin{pmatrix}0 & \frac{a_1}{1+a_2}\\ -\frac{a_1}{1+a_2} & 0\end{pmatrix};\quad R^2_2=\frac{1}{d}\begin{pmatrix}0 & \frac{a_1}{1+a_2}\\ -\frac{a_1}{1+a_2} & 0\end{pmatrix}.
\end{aligned} \qquad (6.108)$$

□

We have to complete the approximation algorithms, presented above, with the initial conditions. The initial situation is characterized by the probability distribution of q_0, which is given by the fixed point q_{-1}, followed by a transition equation. We distinguish two situations.

- $q_{-1} \in M$. Then, q_{-1} determines the manifold $N(q_0) = N(q_{-1})$, on which the coefficients move. There is no problem in initiating the algorithm, because the initial values of the vector fields F_j, Z_l, W_{jl} in the point q_{-1}, i.e. the values of $\psi(\kappa_{-1},\zeta_{-1}), F_i(\kappa_{-1},\zeta_{-1})$, $Z_l(\kappa_{-1},\zeta_{-1})$ and $W_{jl}(\kappa_{-1},\zeta_{-1})$, are well defined and determine the start of the approximation algorithm.

- $q_{-1} \notin M$. For instance, we encountered this situation in example 3.10. The vector fields F_j, Z_l, W_{jl} are not well defined in q_{-1}. Yet, it is well possible to indicate how to choose initial conditions for these vector fields, as we shall do below. We restrict ourselves to the Geodesic and Geodesic plus Noise Models, defined at the end of chapter five. The manifold M and the geodesic tangent field F are as defined in example 3.10. Note that we suppose that the manifold satisfies the separation criterion, see end of chapter five. According to corollary 3.15 in chapter three, there is a chart $(\tilde{U}_\beta, x_\beta)$ of M, such that $F_1 = F; F_2 = c\frac{\partial}{\partial r}$ and $Z_l \perp F_i$ for all l, i. As

$$0 = F_j\langle F_i, Z_l\rangle = \langle F_i, \nabla_{F_j} Z_l\rangle$$

also $W_{jl} \perp F_i$ for all j, l, i. In case of the Geodesic plus Noise-Model, we have $W_{2l} = 0$ everywhere, because of the separation criterion. The value $\psi(\kappa_{-1}, \zeta_{-1}) = (a_1(\kappa_{-1}, \zeta_{-1}), \cdots, a_n(\kappa_{-1}, \zeta_{-1}), r(\kappa_{-1}, \zeta_{-1}))$ and the log-noise-vector field F_2 are well defined in q_{-1}, but the other fields $F = F_1$, Z_l and W_{1l} are not. We take limits. It is possible to show that we can take the chart $(\tilde{U}_\beta = M, x_\beta)$, such that

$$Z_l(q_{-1}) = \lim_{\kappa \to \kappa_{-1}} Z_l(\kappa, \zeta) = 0,$$

see the definition of the "normal" chart $(\tilde{U}_\beta, x_\beta)$ in example 3.10, continuation, in chapter three. The value of the vector field

$$F_1(q_{-1}) := \lim_{\kappa \to \kappa_{-1}} F_1(\kappa, \zeta)$$

determines the direction of the geodesic, and can be considered as a hyperparameter of the Model. The values of

$$W_{1l}(q_{-1}) := \lim_{\kappa \to \kappa_{-1}} W_{1l}(\kappa, \zeta)$$

have to be taken perpendicular (in the sense of the metric) to $F_1(q_{-1})$ and have to be independent; for the rest these are free. As the definition of the normal chart in example 3.10 shows, the values of $W_{1l}(q_{-1})$ only determine a particular chart $(\tilde{U}_\beta, x_\beta)$ with the above mentioned properties. Other values would give another chart with the same properties.

6.3 Identification of hyperparameters

The class of processes that can be described by autoregressive models with time-varying coefficients is much wider than the class of processes generated by standard (constant-coefficient) AR models. Of course, there is a price to be paid for this descriptive power, and part of that price lies in the number of hyperparameters that have to be specified. In this section, we shall discuss methods to determine the hyperparameters from data in a number of special cases, corresponding to relatively simple types of time-varying autoregressive models. Our methods will be approximative and partially heuristic. However, the methods, to be presented in subsection 6.3.1, below have been subjected to extensive numerical testing and we shall report on the experience gained from this.

6.3.1 Geodesic Model on $M \subset \frac{1}{\sigma}AR_n$

The Geodesic Model was defined in chapter three, section 3.8. It is a zero-curvature model. We shall treat a model, where the measurement noise level σ^2 is considered a constant, and with direction space $\mathcal{L} = \boldsymbol{span}\{F\}$, where F is the tangent field of geodesics through a point $q_{-1} \in \frac{1}{\sigma}AR_n$. We shall formally work with a chart (U_β, x_β) as constructed in chapter three, lemma 3.14, but we only use the fact that such a chart exists; we are not assuming that we have an analytical expression for the chart. We only assume that we have analytical expressions for the Christoffel- and curvature symbols on the natural chart. On chart (U_β, x_β), the transition equation is linear. The submanifold $N(q_0)$, i.e. the geodesic through q_{-1} on which the coefficients move, is given on this chart by:

$$N(q_0) = \{q \in U_\beta | q = \frac{1}{\sigma}(x^n - a_1(\kappa,\zeta)x^{n-1} + \cdots - a_n(\kappa,\zeta)) \text{ with } \zeta = \zeta_0\},$$

for some vector $\zeta_0 \in \mathbb{R}^{n-1}$, just as we have seen in subsection 6.1.2.

Estimation of hyperparameters is based on maximization of an *approximate* likelihood. We do not compute the likelihood of the exact model of the Geodesic Model:

$$\begin{aligned} \kappa_0 &\stackrel{d}{=} \mathcal{N}(\kappa_{-1}, \tau^2 Q^2) \\ \kappa_{t+1} &= \kappa_t + a\theta_t + \tau\lambda_t \\ \theta_{t+1} &= a\theta_t + \tau\lambda_t \\ y_t &= \phi_t' a(\kappa_t, \zeta_0) + \sigma_t \epsilon_t, \end{aligned} \tag{6.109}$$

but we linearize the model in a point $\kappa^p = (\kappa_0^p, \cdots, \kappa_T^p)$, which will be specified later. Then we obtain the model:

$$\begin{aligned} \mathcal{A} &= \begin{pmatrix} 1 & a \\ 0 & a \end{pmatrix}; \quad \mathcal{B} = \begin{pmatrix} \tau \\ \tau \end{pmatrix}; \quad C_t = \begin{pmatrix} \phi_t' \frac{\partial a}{\partial \kappa}(\kappa_t^p, \zeta_0) & 0 \end{pmatrix}; \quad (\phi_t = \begin{pmatrix} y_{t-1} \\ \vdots \\ y_{t-n} \end{pmatrix}) \\ x_t &= \begin{pmatrix} \kappa_t \\ \theta_t \end{pmatrix} \\ z_t &= y_t - \phi_t' a(\kappa_t^p, \zeta_0) + \phi_t' \frac{\partial a}{\partial \kappa}(\kappa_t^p, \zeta_0)\kappa_t^p \end{aligned} \tag{6.110}$$

$$\begin{aligned} x_{t+1} &= \mathcal{A}x_t + \mathcal{B}\lambda_{t-1} \\ z_t &= C_t x_t + \sigma\epsilon_t \\ x_0 &\stackrel{d}{=} \mathcal{N}(\begin{pmatrix} \kappa_{-1} \\ 0 \end{pmatrix}, \mathcal{B}Q^2\mathcal{B}'). \end{aligned} \tag{6.111}$$

In the first instance, we shall restrict ourselves to Special Geodesic Models, i.e. we assume that $a = 0$. We shall give a method to determine the following hyperparameters:

- the noise-level σ^2.

- the hyperparameter ζ_0 that determines the geodesic $N(q_0)$. This hyperparameter is crucial. Wrong specification makes all coefficient estimates go in the wrong way, hence it yields bad results.
  In the original Geodesic Model, ζ_0 was just a parameter with a prior distribution, determined by the prior distribution of q_0. In subsection 6.1.2, we fixed it and made a hyperparameter out of it. It has been our experience that it is easier in identification to treat ζ_0 as hyperparameter than as ordinary parameter. We refer to this issue in the next subsection 6.3.2.

- ratio of the measurement noise level σ^2 versus the tensor eigenvalue τ^2 (or system noise variance) $NVR = \frac{\tau^2}{\sigma^2}$.

By applying the Kalman Filter to model 6.111, we can calculate $\kappa_{t|t-1}$, the conditional expectation of κ_t given $z_0, \cdots, z_{t-1}$, according to model (6.111). At the moment of computing $\kappa_{t|t-1}$, we only need the values of the past components $\kappa_0^p, \cdots, \kappa_{t-1}^p$ of the point κ^p in which we linearize model 6.109. Hence, we can define the point κ^p by the Kalman Filter itself by putting κ_t^p equal to $\kappa_{t|t-1}$. In doing so, we are just applying the Extended Kalman Filter to model 6.109, see Anderson and Moore (1979), chapter 8.
We shall use the likelihood of the linearized model. The approximate log-likelihood can be obtained by Prediction Error Decomposition, where the prediction errors are produced by the Kalman Filter, applied to this model. The method of Prediction Error Decomposition is described in Harvey (1990), chapter 3. We shall elaborate the method further below. Briefly, a procedure to estimate the hyperparameters, that worked well in our experiments, had the following general outline.
One can concentrate the noise-level out of Kalman Filter and log-likelihood. This leaves the hyperparameters ζ_0 and NVR to be determined. In the experiments, we performed with this model, the fact is used that

$$\min_{\zeta_0, NVR} l_c(y_0, \cdots, y_T, \zeta_0, NVR) = \min_{NVR}(\min_{\zeta_0} l_c(y_0, \cdots, y_T, \zeta_0, NVR)), \tag{6.112}$$

where l_c is minus the concentrated approximate log-likelihood.
Minimization w.r.t. ζ_0 can be carried out using Gauss-Newton iteration. Minimization w.r.t. the NVR has been done using a discrete parameter search method.

We shall elaborate the procedure now. The Kalman Filter for the model (6.111) can be given the following form: According to this model, the prediction error

$$pe_t = z_t - C_t x_{t|t-1} = y_t - \phi_t' a(\kappa_{t|t-1}, \zeta_0) \tag{6.113}$$

is normally distributed, conditional on the past $z_0, \cdots, z_{t-1}$ with mean zero and variance

$$pev_t = C_t \Sigma_{t|t-1} C_t' + \sigma^2 = \phi_t' \frac{\partial a}{\partial \kappa}(\kappa_{t|t-1}, \zeta_0) \mathcal{V}_{t|t-1} \frac{\partial a'}{\partial \kappa}(\kappa_{t|t-1}, \zeta_0) \phi_t + \sigma^2. \tag{6.114}$$

We define the "normalized prediction error at time t," or npe_t, by:

$$npe_t = \frac{pe_t}{\sqrt{pev_t}}. \tag{6.115}$$

The normalized prediction error at time t has a variance equal to one, conditionally on the previous data $y_0, \cdots, y_{t-1}$. The (modified) gain vector K_t is defined by:

$$K_t = \frac{\Sigma_{t|t-1} C_t'}{\sqrt{C_t \Sigma_{t|t-1} C_t' + \sigma^2}} = \frac{\Sigma_{t|t-1} \begin{pmatrix} \frac{\partial a}{\partial \kappa}'(\kappa_{t|t-1}, \zeta_0)\phi_t \\ 0 \end{pmatrix}}{\sqrt{pev_t}}. \tag{6.116}$$

Normalized prediction error and (modified) gain vector are used in the update equations:

$$\begin{aligned} x_{t+1|t} &= \mathcal{A} x_{t|t} \\ x_{t|t} &= x_{t|t-1} + K_t npe_t \\ \Sigma_{t+1|t} &= \mathcal{A}\Sigma_{t|t-1}\mathcal{A}' - \mathcal{A}K_tK_t'\mathcal{A}' + \mathcal{B}\mathcal{B}'. \end{aligned} \tag{6.117}$$

Then, after introducing

$$\begin{aligned} n\tilde{p}e_t &= \sigma\, npe_t \quad \tilde{K}_t = \frac{K_t}{\sigma} \quad \tilde{\Sigma}_{t|t-1} = \frac{\Sigma_{t|t-1}}{\sigma^2} \quad p\tilde{e}v_t = \frac{pev_t}{\sigma^2} \\ \mathcal{N} &= \frac{\mathcal{B}}{\sigma} \quad \text{(contains the square root of the } NVR\text{)}, \end{aligned} \tag{6.118}$$

the Kalman Filter algorithm becomes:

initialization

$$x_{0|-1} = \begin{pmatrix} \kappa_{-1} \\ 0 \end{pmatrix}; \quad \tilde{\Sigma}_{0|-1} = \mathcal{N}Q^2\mathcal{N}'$$

recursion

$$pe_t = y_t - \phi_t' a(\kappa_{t|t-1}, \zeta_0) \tag{6.119}$$

$$C_t = \begin{pmatrix} \phi_t' \frac{\partial a}{\partial \kappa}(\kappa_{t|t-1}, \zeta_0) & 0 \end{pmatrix} \tag{6.120}$$

$$\begin{aligned} p\tilde{e}v_t &= C_t \tilde{\Sigma}_{t|t-1} C_t' + 1 \\ \tilde{K}_t &= \frac{\tilde{\Sigma}_{t|t-1} C_t'}{\sqrt{p\tilde{e}v_t}} \\ n\tilde{p}e_t &= \frac{pe_t}{\sqrt{p\tilde{e}v_t}} \end{aligned} \tag{6.121}$$

updates

$$\begin{aligned} x_{t|t} &= x_{t|t-1} + \tilde{K}_t n\tilde{p}e_t \\ x_{t+1|t} &= \mathcal{A} x_{t|t} \qquad \kappa_{t|s} = (x_{t|s})_1 \\ \tilde{\Sigma}_{t+1|t} &= \mathcal{A}\tilde{\Sigma}_{t|t-1}\mathcal{A}' - \mathcal{A}\tilde{K}_t\tilde{K}_t'\mathcal{A}' + \mathcal{N}\mathcal{N}'. \end{aligned} \tag{6.122}$$

In these equations, σ^2 has been eliminated. Next, we shall explain the well known (Harvey,1990) method of concentrating the noise-level σ^2 out of the approximate likelihood. We define $L(y_0, \cdots, y_T)$ as:

$$L(y_0, \cdots, y_T) = -\frac{2}{T+1} \log p(y_0, \cdots, y_T) = -\frac{2}{T+1} \sum_{t=0}^{T} \log p(y_t | y_{t-1}, \cdots, y_0), \tag{6.123}$$

where p is that approximate likelihood. The likelihood of y_t, conditional on $y_{t-1}, \cdots, y_0$, according to model (6.111), is implied by the conditional distribution of npe_t, given above, hence:

$$\begin{aligned} -\log p(y_t|y_{t-1}, \cdots, y_0) &= \frac{1}{2}\log(2\pi) + \frac{1}{2}\log(pev_t) + \frac{1}{2}npe_t^2 \\ &= \frac{1}{2}(\log(2\pi) + \log(\sigma^2) + \log(p\tilde{e}v_t) + \frac{n\tilde{p}e_t^2}{\sigma^2}). \end{aligned} \quad (6.124)$$

As $p\tilde{e}v_t$ and $n\tilde{p}e_t$ don't depend on σ^2, we have:

$$\frac{\partial}{\partial\sigma^2}L(y_0, \cdots, y_T) = -\frac{1}{\sigma^2} + \frac{1}{T+1}\sum_{t=0}^{T} n\tilde{p}e_t^2,$$

and we can express the value of the noise level $\hat{\sigma}^2$ that minimizes the function L in the other terms:

$$\hat{\sigma}^2 = \frac{1}{T+1}\sum_{t=0}^{T} n\tilde{p}e_t^2. \quad (6.125)$$

By substituting this expression in L (and subtracting $\log(2\pi) + 1$), we obtain the *concentrated approximate* $-\frac{2}{T+1}\log$ *likelihood* $L_c(y_0, \cdots, y_T)$:

$$L_c(y_0, \cdots, y_T) = \log(\frac{1}{T+1}\sum_{t=0}^{T} n\tilde{p}e_t^2) + \frac{1}{T+1}\sum_{t=0}^{T}\log(p\tilde{e}v_t). \quad (6.126)$$

The approximate likelihood has a maximum in the point $(\bar{\zeta}_0, N\bar{V}R, \bar{\sigma}^2)$ if and only if L_c has a minimum in $(\bar{\zeta}_0, N\bar{V}R)$, and $\bar{\sigma}^2 = \frac{1}{T+1}\sum_{t=0}^{T} n\tilde{p}e_t^2$. This does not mean that always an error in the location $(\bar{\zeta}_0, N\bar{V}R)$ of the minimum of L_c would yield a comparable error in the location of the maximum of the approximate likelihood. Harvey (1990, p.183) reports that there may arise problems if the value, found for $\bar{\sigma}^2$, is very close to zero. Yet, concentrating the noise-level out is an computationally efficient way to calculate the other parameter estimates, because it reduces the dimension of the problem.
We continue with the function L_c, and show that minimization of this function w.r.t. the other parameters ζ_0 and NVR is just a non-linear least squares problem. Note that

$$(T+1)\exp(L_c) = n'n \quad \text{where } n = \exp(\frac{\frac{1}{2}}{T+1}\sum_{t=0}^{T}\log(p\tilde{e}v_t))\begin{pmatrix} n\tilde{p}e_0 \\ \vdots \\ n\tilde{p}e_T \end{pmatrix}. \quad (6.127)$$

We explain the Gauss-Newton method, we used for minimization of L_c w.r.t. ζ_0. This method is given by the iterations

$$\hat{\zeta}_0^{p+1} = \hat{\zeta}_0^{p} - \gamma(p)(\mathrm{D}n(\hat{\zeta}_0^{p})'\mathrm{D}n(\hat{\zeta}_0^{p}))^{-1}\mathrm{D}n(\hat{\zeta}_0^{p})'n(\hat{\zeta}_0^{p}). \quad (6.128)$$

Here, $\mathrm{D}n$ is the derivative of n w.r.t. ζ_0, and $\gamma(p)$ is some gain constant to modify the size of the iteration step. In the strict sense of Gauss-Newton algorithm, $\gamma(p)$ should be equal to one. Accordingly, it is only necessary to compute the derivative $\mathrm{D}n$ in order to

apply this algorithm. This can be incorporated in the recursions of the Kalman Filter by differentiating the equations (6.122)w.r.t. ζ_0. We indicate the derivative of a certain function f with respect to (a component of) ζ_0 by $\mathrm{D}f$, then differentiating equations (6.122) and (6.127) yields:

initialization

$$\mathrm{D}x_{0|-1} = 0; \quad \mathrm{D}\tilde{\Sigma}_{0|-1} = 0$$

recursion

$$\mathrm{D}pe = -\phi_t' \frac{\partial a}{\partial \zeta}(\kappa_{t|t-1}, \zeta) - \phi_t' \frac{\partial a}{\partial \kappa}(\kappa_{t|t-1}, \zeta)\mathrm{D}\kappa_{t|t-1} \tag{6.129}$$

$$\mathrm{D}C_t = \begin{pmatrix} \phi_t' \frac{\partial^2 a}{\partial \zeta \partial \kappa}(\kappa_{t|t-1}, \zeta) + \phi_t' \frac{\partial^2 a}{\partial \kappa^2}(\kappa_{t|t-1}, \zeta)\mathrm{D}\kappa_{t|t-1} & 0 \end{pmatrix} \tag{6.130}$$

$$\mathrm{D}p\tilde{e}v_t = \mathrm{D}C_t\tilde{\Sigma}_{t|t-1}C_t' + C_t\mathrm{D}\tilde{\Sigma}_{t|t-1}C_t' + C_t\tilde{\Sigma}_{t|t-1}\mathrm{D}C_t'$$

$$\mathrm{D}\tilde{K}_t = \frac{\mathrm{D}\tilde{\Sigma}_{t|t-1}C_t'p\tilde{e}v_t + \tilde{\Sigma}_{t|t-1}\mathrm{D}C_t'p\tilde{e}v_t - \frac{1}{2}\tilde{\Sigma}_{t|t-1}C_t'\mathrm{D}p\tilde{e}v_t}{p\tilde{e}v_t\sqrt{p\tilde{e}v_t}}$$

$$\mathrm{D}n\tilde{p}e_t = \frac{\mathrm{D}pe_tp\tilde{e}v_t - \frac{1}{2}pe_t\mathrm{D}p\tilde{e}v_t}{p\tilde{e}v_t\sqrt{p\tilde{e}v_t}}$$

updates

$$\mathrm{D}x_{t|t} = \mathrm{D}x_{t|t-1} + \mathrm{D}\tilde{K}_tn\tilde{p}e_t + \tilde{K}_t\mathrm{D}n\tilde{p}e_t$$

$$\mathrm{D}x_{t+1|t} = \mathcal{A}\mathrm{D}x_{t|t} \qquad \mathrm{D}\kappa_{t|s} = (\mathrm{D}x_{t|s})_1$$

$$\mathrm{D}\tilde{\Sigma}_{t+1|t} = \mathcal{A}\mathrm{D}\tilde{\Sigma}_{t|t-1}\mathcal{A}' - \mathcal{A}\mathrm{D}\tilde{K}_t\tilde{K}_t'\mathcal{A}' - \mathcal{A}\tilde{K}_t\mathrm{D}\tilde{K}_t'\mathcal{A}'$$

derivative of n (6.131)

$$\mathrm{D}n = (\tfrac{\frac{1}{2}}{T+1}\sum_{t=0}^{T}\frac{\mathrm{D}p\tilde{e}v_t}{p\tilde{e}v_t})n + \exp(\tfrac{\frac{1}{2}}{T+1}\sum_{t=0}^{T}\log(p\tilde{e}v_t))\begin{pmatrix} \mathrm{D}n\tilde{p}e_0 \\ \vdots \\ \mathrm{D}n\tilde{p}e_t \end{pmatrix}.$$

If no explicit analytic formula for $a(\kappa, \zeta)$ is available, we can use the approximation method of the previous section 6.2 in equations (6.119, 6.120, 6.129) and (6.130). Accordingly, one Gauss-Newton-ζ_0 search step from $\hat{\zeta}_0^{\,p}$ to $\hat{\zeta}_0^{\,p+1}$ consists of:

1. the Extended Kalman Filter to obtain the estimates $\kappa_{t|t-1}$, given the parameter value $\zeta = \hat{\zeta}_0^{\,p}$, from algorithm (6.122) in combination with the approximation algorithm (6.101-6.107);

2. the construction of n and the criterion $n'n$ which has to be minimized in equation (6.127);

3. building up the derivative $\mathrm{D}n$ of n from algorithm (6.131) in combination with the approximation algorithm (6.101-6.107) at the same time with the E.K.F.;

4. the update $\hat{\zeta}_0^{\,p} \mapsto \hat{\zeta}_0^{\,p+1}$ from equation (6.128).

In the approximation algorithm (6.101-6.107), all terms containing $\hat{\zeta}_{t+1} - \hat{\zeta}_t$ are canceled, because during one Gauss-Newton-ζ_0-search step 1,2,3,4 all $\hat{\zeta}_t$ are equal to $\hat{\zeta}_0^{\,p}$ for all t.

Finally, we used discrete parameter search in order to find a minimizing NVR -value for the function $NVR \mapsto \min_{\zeta_0} L_c(\zeta_0, NVR)$. The method is laborious, but worked fine in the experiments that we performed. At every step q in this search (when the estimate of the NVR is $N\hat{V}R^q$) a full Gauss-Newton iteration to obtain a reasonable approximation of the minimizing $\hat{\zeta}_0$ of $L_c(\zeta_0, N\hat{V}R^q)$ was executed. Of course, for non-complete metrics, this is only feasible for small values of the NVR; if not the estimates $\kappa_{t|t-1}$ will rapidly run out of the domain V_β. The experimental results with the Gauss-Newton method showed that convergence success of the estimates $(\hat{\zeta}_0^{\,p})_{p\in\mathbb{N}}$ was sometimes dependent on a well chosen initial value (in combination with a well chosen estimate of the NVR). This indicates that a full Newton procedure would be advisable, but second derivatives w.r.t. ζ_0 are not available with the approximation algorithm of the last section.
In these experiments, autoregressive processes with coefficients on $\frac{1}{\sigma}AR_2$ were generated, according to the Geodesic Model, and afterwards analyzed by the above described method to determine the three hyperparameters. When the estimate of the NVR was close to the true one, then the Gauss-Newton procedure for $\hat{\zeta}_0$ worked reasonably well. See some experimental results in the next section.

Remark 6.8 One can replace $\kappa_{t|t-1}$ by $\hat{\kappa}_t^t$, i.e. the prediction of κ_t by Fahrmeir/Kaufmann's method. The estimates $\kappa_{t-1|t-1}$ of the E.K.F. and $\hat{\kappa}_{t-1}^t$ of Fahrmeir/Kaufmann's method are both approximations of the $(t-1)$th component of the mode of the posterior density $p(\kappa_0, \cdots, \kappa_{t-1}|y_0, \cdots, y_{t-1})$. The transition from filtered estimate $\hat{\kappa}_{t-1}^t$ to prediction $\hat{\kappa}_t^t$ follows the same linear rule as the transition $\kappa_{t-1|t-1} \mapsto \kappa_{t|t-1}$. For a linear model the mode coincides with the conditional expectation.

Remark 6.9 The Gauss-Newton procedure to determine $\hat{\zeta}_0$, presented above, has been inspired by the Recursive Prediction Error method of Ljung and Söderström (1983). That latter method adapts the estimate of the hyperparameter *on-line*, i.e. at every time instant t, whereas the first is *off-line*, i.e. the procedure uses all the $T+1$ observations in every iteration. One can make the method more on line by using the philosophy of the Fahrmeir/Kaufmann's method. In fact, it is possible to incorporate one iteration of the Gauss-Newton procedure, presented above, in every five-step-stage of Fahrmeir/Kaufmann's method.

6.3.2 Geodesic plus Noise Model

The Geodesic plus Noise model was defined in chapter five, section 5.3. It is a zero-curvature model, where the coefficients move on a two-dimensional submanifold $N(q_0)$. The coefficients $\vec{a}_t$ move on a geodesic on AR_n, and the log-noise-levels $r_t = \log(\sigma_t)$ are also time-varying. On a suitable chart (U_β, x_β) the transition equation will be linear,

and there will be a vector $\zeta_0 \in \mathbb{R}^{n-1}$, such that the manifold $N(q_0)$ can be written as:

$$N(q_0) = \{q \in U_\beta | q = e^{-\mu\rho}(x^n - a_1(\kappa,\zeta)x^{n-1} + \cdots - a_n(\kappa,\zeta) \text{ with } \zeta = \zeta_0\}.$$

The exact system of equations for a Geodesic plus Noise model on such a chart (U_β, x_β) is:

$$\begin{aligned}
\tilde{\kappa}_0 &\stackrel{d}{=} \mathcal{N}(\tilde{\kappa}_{-1}, \tau_1^2 Q^2) \\
\rho_0 &\stackrel{d}{=} \mathcal{N}(\rho_{-1}, \tau_2^2 q^2) \qquad (6.132) \\
\tilde{\kappa}_{t+1} &= \tilde{\kappa}_t + A_1\theta_t + \tau_1\lambda_t \\
\theta_{t+1} &= A_1\theta_t + \tau_1\lambda_t \\
\rho_{t+1} &= \rho_t + A_2\vartheta_t + \tau_2\omega_t \\
\vartheta_{t+1} &= A_2\vartheta_t + \tau_2\omega_t \\
y_t &= \phi_t' a(\tilde{\kappa}_t, \zeta_0) + e^{\mu\rho_t}\epsilon_t, \qquad (6.133)
\end{aligned}$$

where $(\lambda_t)_{t\in\mathbf{Z}}$, $(\omega_t)_{t\in\mathbf{Z}}$ and $(\epsilon_t)_{t\in\mathbf{Z}}$ are independent sequences of unit-variance Gaussian white noise.

We are going to propose a method for estimating the hyperparameters A_1, τ_1, A_2, τ_2 of the transition equations and ζ_0 of the measurement equation. The method will be partially heuristic, because we shall make a number of approximations without evaluating the error of approximation. For the hyperparameters of the transition equation, it is interesting to use the Expectation-Maximization method in order to maximize the likelihood. It is then convenient to look at ζ_0 not as hyperparameter, but as an ordinary parameter with a prior distribution. To estimate ζ_0, we maximize the posterior density

$$p(\zeta_0, \tilde{\kappa}_0, \cdots, \tilde{\kappa}_T, \rho_0, \cdots, \rho_T | y_0, \cdots, y_T).$$

We do this in the following order:

$$\max_{\zeta_0,\tilde{\kappa},\rho} p(\cdot|y_0,\cdots,y_T) = \max_{\rho}(\max_{\zeta_0,\tilde{\kappa}} p(\cdot|y_0,\cdots,y_T)).$$

One of the approximations that we shall make is that we replace the maximum a posteriori estimate for ζ_0 by a maximum likelihood estimate. We shortly describe what is happening then.

Remark 6.10 We can extend the priors for the coefficient coordinates $\kappa = (\tilde{\kappa}, \rho)$ with a prior for the other coordinates, ζ_0. This means that the ζ-coordinates of the intial point q_0 are not considered as a hyperparameter. We add the equation

$$\zeta_0 \stackrel{d}{=} \mathcal{N}(\zeta_{-1}, Q_z^2) \qquad (6.134)$$

to the equations (6.132). The prior for ζ_0 is considered to be independent of the other coefficient coordinate priors. One can try and maximize now the posterior density with respect to $\tilde{\kappa}, \rho$ ànd ζ_0. Then, one obtains a *maximum a posterior estimate* of ζ_0. It is difficult to implement the method of Fahrmeir and Kaufmann, because the gain matrix

U of the Gauss-Newton iteration is not pentadiagonal anymore, but has non-zero elements everywhere in the last column and last row. Note that the situation is non-linear even if the model is based on the Euclidean metric. But, if we were able to compute this maximum a posteriori estimate of ζ_0, would the answer be very different from the maximum likelihood estimate of ζ_0 without putting a prior on ζ_0? We compare the two estimates.

The notations $l(y|\zeta_0) = l(y_0, \cdots, y_T|\zeta_0)$ and $p(y|\zeta_0) = p(y_0, \cdots, y_T|\zeta_0)$ indicate the log-likelihood and likelihood, respectively, according to the Geodesic plus Noise Model with hyperparameter ζ_0. We denote the conditional expectation of a stochastic variable z, according to this model given all data $y_0, \cdots, y_T$, by $\boldsymbol{E}^{\mathcal{Y}}_{\zeta_0} z$. To calculate the maximum likelihood estimate, one has to resolve the equation

$$0 = \frac{\partial}{\partial \zeta_0} p(y|\zeta_0).$$

The right hand side can be rewritten as follows:

$$\begin{aligned} \frac{\partial}{\partial \zeta_0} p(y|\zeta_0) &= \int \frac{\partial}{\partial \zeta_0} p(y|\kappa, \zeta_0) p(\kappa) d\kappa = \\ &= \int \frac{\partial}{\partial \zeta_0} l(y|\kappa, \zeta_0) p(y, \kappa|\zeta_0) d\kappa = \\ &= \int \frac{\partial}{\partial \zeta_0} l(y|\kappa, \zeta_0) p(\kappa|y, \zeta_0) d\kappa \; p(y|\zeta_0) \end{aligned}$$

hence:

$$\frac{\partial}{\partial \zeta_0} l(y|\zeta_0) = \boldsymbol{E}^{\mathcal{Y}}_{\zeta_0} \frac{\partial}{\partial \zeta_0} l(y|\kappa, \zeta_0) = \boldsymbol{E}^{\mathcal{Y}}_{\zeta_0} \frac{\partial}{\partial \zeta_0} l(y, \zeta_0|\kappa) - \frac{\partial}{\partial \zeta_0} l(\zeta_0), \tag{6.135}$$

because

$$\frac{\partial}{\partial \zeta_0} l(y, \zeta_0|\kappa) = \frac{\partial}{\partial \zeta_0} l(\zeta_0) + \frac{\partial}{\partial \zeta_0} l(y|\kappa, \zeta_0), \tag{6.136}$$

as ζ_0 is independent of $\kappa = (\tilde{\kappa}, \rho)$. This implies that computation of the maximum likelihood estimate requires the solution of

$$\boldsymbol{E}^{\mathcal{Y}}_{\zeta_0} \frac{\partial}{\partial \zeta_0} l(y, \zeta_0|\kappa) = \frac{\partial}{\partial \zeta_0} l(\zeta_0), \tag{6.137}$$

or, taking the mean over $T+1$ observations,

$$\boldsymbol{E}^{\mathcal{Y}}_{\zeta_0} \frac{\partial}{\partial \zeta_0} \frac{1}{T+1} l(y, \zeta_0|\kappa) = \frac{1}{T+1} \frac{\partial}{\partial \zeta_0} l(\zeta_0), \tag{6.138}$$

i.e.

$$\frac{1}{T+1} \sum_{t=0}^{T} \boldsymbol{E}^{\mathcal{Y}}_{\zeta_0} (e^{-2r(\rho_t)} \frac{\partial}{\partial \zeta_0} (y_t - a(\tilde{\kappa}_t, \zeta_0)' \phi_t)^2 = \frac{1}{T+1} \frac{\partial}{\partial \zeta_0} (\zeta_0 - \zeta_{-1})' Q_z^{-2} (\zeta_0 - \zeta_{-1}). \tag{6.139}$$

On the other hand, to compute the joint maximum a posteriori estimate of $\zeta_0, \kappa_0, \cdots, \kappa_T$ one has to solve the equations:

$$\frac{\partial}{\partial \kappa_0}\frac{1}{T+1}l(y,\kappa,\zeta_0) = 0 \tag{6.140}$$

$$\vdots \quad \vdots$$

$$\frac{\partial}{\partial \kappa_T}\frac{1}{T+1}l(y,\kappa,\zeta_0) = 0 \tag{6.141}$$

$$\frac{\partial}{\partial \zeta_0}\frac{1}{T+1}l(y,\kappa,\zeta_0) = 0, \text{ i.e.} \tag{6.142}$$

$$\frac{1}{T+1}\sum_{t=0}^{T} e^{-2r(\rho_t)}\frac{\partial}{\partial \zeta_0}(y_t - a(\tilde{\kappa}_t,\zeta_0)'\phi_t)^2 = 0 \tag{6.143}$$

We compare the last equation (6.143) with (6.139). The right hand side of (6.139) will be close to zero if the prior variance Q_z^2 of ζ_0 or the number of data $T+1$ is large. The difference between the left hand side of both equations (6.143) and (6.139) is that in the latter a conditional expectation given all data is taken of $\frac{\partial}{\partial \zeta_0}\frac{1}{T+1}l(y,\zeta_0|\kappa)$, whereas in the first equation the coordinates of a stationary point (κ,ζ_0) are inserted. □

Now, we elaborate the proposed hyperparameter identification procedure.
An alternative to discrete parameter search in maximizing the likelihood is the Expectation-Maximization algorithm. Fahrmeir (1992) proposes this algorithm for determining the hyperparameters of the transition equation in Dynamic Exponential Family Regression. Shumway and Stoffer (1982) use the algorithm to estimate the hyperparameters of a linear dynamic stochastic system. We shortly describe the backgrounds of the EM algorithm. We refer to Harvey(1989), Dempster, Laird, Rubin (1977) and especially Wu (1983). Let ω be the vector of the afore-mentioned hyperparameters of the transition equation:

$$\omega = \begin{pmatrix} \tau_1^2 \\ \tau_2^2 \\ A_1 \\ A_2 \end{pmatrix}.$$

We denote the log-likelihood of the Geodesic plus Noise model with hyperparameters ω by l_ω :

$$l_\omega(y_0,\cdots,y_T) = \log p_\omega(y_0,\cdots,y_T).$$

By $\boldsymbol{E}_\omega^{\boldsymbol{y}} z$, we denote the expectation of a stochastic variable z given $y_0,\cdots,y_T$ under the Geodesic plus Noise model with hyperparameters ω. Then we have:

$$l_{\omega'}(y) = \boldsymbol{E}_\omega^{\boldsymbol{y}} l_{\omega'}(y) = \boldsymbol{E}_\omega^{\boldsymbol{y}} l_{\omega'}(y,\tilde{\kappa},\rho) - \boldsymbol{E}_\omega^{\boldsymbol{y}} l_{\omega'}(\tilde{\kappa},\rho|y) = Q(\omega'|\omega) - H(\omega'|\omega).$$

Furthermore,

$$H(\omega'|\omega) - H(\omega|\omega) = \boldsymbol{E}_\omega^{\boldsymbol{y}} \log\left(\frac{p_{\omega'}(\tilde{\kappa},\rho|y)}{p_\omega(\tilde{\kappa},\rho|y)}\right) \leq \log\left(\boldsymbol{E}_\omega^{\boldsymbol{y}} \frac{p_{\omega'}(\tilde{\kappa},\rho|y)}{p_\omega(\tilde{\kappa},\rho|y)}\right) = 0.$$

This shows that for any ω' that, given ω, maximizes the function

$$\omega' \mapsto Q(\omega'|\omega) = \boldsymbol{E}_{\omega}^{\mathcal{Y}} l_{\omega'}(y, \tilde{\kappa}, \rho),$$

we have:

$$l_{\omega'}(y) \geq l_{\omega}(y).$$

The EM-algorithm defines a sequence $(\omega_n)_{n\in\mathbf{N}}$ by:

$$\omega_{n+1} \text{ maximizes } \omega' \mapsto Q(\omega'|\omega_n) \text{ over the parameter space.} \tag{6.144}$$

Accordingly, the corresponding sequence of log-likelihoods $(l_{\omega_n})_{n\in\mathbf{N}}$ increases. The parameter space Ω is given by:

$$\Omega = [0,\infty) \times [0,\infty) \times \mathbb{R} \times \mathbb{R} \text{ or } \Omega = [0,\infty) \times [0,\infty) \times [0,\infty) \times [0,\infty).$$

The function $\omega \mapsto l_{\omega}(y)$ is continuous in Ω, and differentiable in the interior of Ω. Also $\omega' \mapsto Q(\omega'|\omega)$; $\omega \mapsto Q(\omega'|\omega)$ are continuous functions for all $\omega', \omega \in \Omega$. It is more difficult to understand, whether the set

$$\Omega_{\omega_0} = \{\omega \in \Omega | l_{\omega}(y) \geq l_{\omega_0}(y)\}$$

is compact for all $\omega_0 \in \Omega$, such that $l_{\omega_0} > -\infty$. Of course, we can restrict the parameter space to a compact subset of Ω with non-empty interior. If Ω_{ω_0} is compact for all $\omega_0 \in \Omega$, then, as Wu proves in his article, the likelihoods $(l_{\omega_n}(y))_{n\in\mathbf{N}}$, corresponding to a sequence $(\omega_n)_{n\in\mathbf{N}}$ according to the EM-algorithm (6.144), converge to a likelihood $l_{\omega_*}(y)$, such that ω_* is a stationary point of $\omega \mapsto l_{\omega}(y)$. The article also gives conditions for this stationary point to be a local maximum. As these are not simple to verify, Wu advises to generate several EM-sequences, starting in different points of the parameter space. It is reported that convergence is slow.

Maximizing the function $\omega' \mapsto Q(\omega'|\omega)$ is the same as maximizing the function

$$\begin{aligned} \omega' \mapsto\ & T\log(\tau_1'^2) + \frac{\sum_{t=0}^{T-1} \boldsymbol{E}_{\omega}^{\mathcal{Y}}(\tilde{\kappa}_{t+1} - \tilde{\kappa}_t - A_1'(\tilde{\kappa}_t - \tilde{\kappa}_{t-1}))^2}{\tau_1'^2} + \\ & + T\log(\tau_2'^2) + \frac{\sum_{t=0}^{T-1} \boldsymbol{E}_{\omega}^{\mathcal{Y}}(\rho_{t+1} - \rho_t - A_2'(\rho_t - \rho_{t+1})^2}{\tau_2'^2}. \end{aligned} \tag{6.145}$$

The nth step of the EM-algorithm yields the following estimates (provided the values of the expressions are in the parameter space Ω):

$$(\hat{A}_1)_{n+1} = \frac{\sum_{t=0}^{T-1} \boldsymbol{E}_{\omega_n}^{\mathcal{Y}}(\tilde{\kappa}_{t+1} - \tilde{\kappa}_t)(\tilde{\kappa}_t - \tilde{\kappa}_{t-1})}{\sum_{t=0}^{T-1} \boldsymbol{E}_{\omega_n}^{\mathcal{Y}}(\tilde{\kappa}_{t+1} - \tilde{\kappa}_t)^2} \tag{6.146}$$

$$(\hat{\tau_1^2})_{n+1} = \frac{1}{T}\sum_{t=0}^{T-1} \boldsymbol{E}_{\omega_n}^{\mathcal{Y}}(\tilde{\kappa}_{t+1} - \tilde{\kappa}_t - (\hat{A}_1)_{n+1}(\tilde{\kappa}_t - \tilde{\kappa}_{t-1}))^2 \tag{6.147}$$

$$(\hat{A}_2)_{n+1} = \frac{\sum_{t=0}^{T-1} \boldsymbol{E}_{\omega_n}^{\mathcal{Y}}(\rho_{t+1} - \rho_t)(\rho_t - \rho_{t-1})}{\sum_{t=0}^{T-1} \boldsymbol{E}_{\omega_n}^{\mathcal{Y}}(\rho_{t+1} - \rho_t)^2} \tag{6.148}$$

$$(\hat{\tau_2^2})_{n+1} = \frac{1}{T}\sum_{t=0}^{T-1} \boldsymbol{E}_{\omega_n}^{\mathcal{Y}}(\rho_{t+1} - \rho_t - (\hat{A}_2)_{n+1}(\rho_t - \rho_{t-1}))^2. \tag{6.149}$$

Next, we are going to make an approximation in the EM-algorithm. We shall replace the conditional expectation, given all data $y_0, \cdots, y_T$ of the coordinates $\tilde{\kappa}_0, \cdots, \tilde{\kappa}_T$, and $\rho_0, \cdots, \rho_T$ by the posterior mode, i.e. by the outcome of Fahrmeir/Kaufmann's method. The conditional variance of $\tilde{\kappa}_0, \cdots, \tilde{\kappa}_T; \rho_0, \cdots, \rho_T$, given all data, will be replaced by the inverse of U, the gain matrix of the Gauss-Newton iteration in Fahrmeir's method. This is the same as approximating the posterior density of the coordinates $\tilde{\kappa}_0, \cdots, \tilde{\kappa}_T; \rho_0, \cdots, \rho_T$ by a Gaussian one. Of course, the approximation in case of $\tilde{\kappa}_0, \cdots, \tilde{\kappa}_T$ makes more sense than in the case of $\rho_0, \cdots, \rho_T$.
In order to resolve the difficulty that Fahrmeir/Kaufmann's method poses, if a maximum a posteriori estimate of ζ_0 is calculated instead of the maximum likelihood estimate, we pursue according to the following practical lines. An alternative to the procedure with, interchangingly, steps to update the coefficients and steps to update the noise levels,is to maximize in the following order:

$$\max_{\zeta_0, \tilde{\kappa}_0, \cdots, \tilde{\kappa}_T, \rho_0, \cdots, \rho_T} l_\omega(y, \zeta_0, \tilde{\kappa}, \rho) = \max_{\rho_0, \cdots, \rho_T} \max_{\zeta_0, \tilde{\kappa}_0, \cdots, \tilde{\kappa}_T} l_\omega(y, \zeta_0, \tilde{\kappa}, \rho),$$

i.e. repeating the coefficient updating step a number of times in order to find the maximizing values of ζ_0 and $\tilde{\kappa}_0, \cdots, \tilde{\kappa}_T$, given $\rho_0, \cdots, \rho_T$, then doing one noise level updating step, then again the coefficient updating step a number of times etc. Between two noise level updating steps a maximum a posteriori estimate of ζ_0 (jointly with $\tilde{\kappa}_0, \cdots, \tilde{\kappa}_T$), conditional on the log-noise levels $\rho_0, \cdots, \rho_T$, is calculated. As we have indicated in the remark above, one can conjecture that there is not much difference between a maximum a posteriori estimate of ζ_0 jointly with $\tilde{\kappa}_0, \cdots, \tilde{\kappa}_T$ and the maximum likelihood estimate of ζ_0. Hence, instead of computing the first type of estimate, one may calculate the latter one for ζ_0 conditional on the log-noise levels. This can be done in the same way as proposed for the geodesic model. The only difference is that the noise levels are not concentrated out of the likelihood in this case: these are treated as fixed. The $-\frac{2}{T+1}$ log-likelihood $L_{\zeta_0}(y_0, \cdots, y_T | \hat{\rho}_0, \cdots, \hat{\rho}_T)$, given the noise levels, gets the form:

$$\begin{aligned}
L_{\zeta_0}(y|\hat{\rho}) &= \log(2\pi) + \frac{1}{T+1}\sum_{t=0}^{T} \log(pev_t) + \frac{1}{T+1}\sum_{t=0}^{T}(upe_t)^2, \\
&\quad \text{with notations as in the previous subsection:} \\
pe_t &= y_t - \phi_t' a(\tilde{\kappa}_{t|t-1}, \zeta_0) \\
pev_t &= \phi_t' \frac{\partial a}{\partial \tilde{\kappa}}(\tilde{\kappa}_{t|t-1}, \zeta_0) \mathcal{V}_{t|t-1} \frac{\partial a'}{\partial \tilde{\kappa}}(\tilde{\kappa}_{t|t-1}, \zeta_0)\phi_t + e^{2\mu\hat{\rho}_t} \\
upe_t &= \frac{pe_t}{\sqrt{pev_t}}.
\end{aligned}$$

Now, using assumption (5.44) that the log-noise levels are bounded from below, we have that there is a positive number c, such that

$$\log(pev_t) \geq \log(c)$$

for all t. Instead of minimizing the function L_{ζ_0} w.r.t. ζ_0, we can equivalently minimize

$$(T+1)(L_{\zeta_0}(y|\hat{\rho}) - \log(c) = n'n$$

$$\text{where } n = \begin{pmatrix} \sqrt{\log(2\pi) + (upe_0)^2 + \log(pev_0) - \log(c)} \\ \vdots \\ \sqrt{\log(2\pi) + (upe_T)^2 + \log(pev_T) - \log(c)} \end{pmatrix}.$$

Hence, the Gauss-Newton method in order to minimize $L_{\zeta_0}(y_0, \cdots, y_T | \hat{\rho}_0, \cdots, \hat{\rho}_T)$ w.r.t. ζ_0 is possible.
The combination of Fahrmeir/Kaufmann's method with the approximate EM-algorithm yields the following procedure:

1. (*Initialization; put* $n = 1$.) Choose $\omega_1 \in \Omega$.

2. (*Recursion; nth step*) Apply Fahrmeir/Kaufmann'method to compute the posterior mode of $\zeta_0, \tilde{\kappa}_0, \cdots, \tilde{\kappa}_T, \rho_0, \cdots, \rho_T$, conditional on the hyperparameters ω_n. Compute the gain matrix U (or, rather, its inverse), corresponding to this posterior mode (step 3 of Fahrmeir's method).

3. Compute ω_{n+1}, using the posterior mode, the inverse of the corresponding gain matrix U and Gaussian approximation from equations (6.146-6.149).

4. Substitute n by $n + 1$; goto step 2.

6.4 Simple Models

In section 3.9 of chapter three, we concluded that the simplest example of a General Model for a time-varying AR(n) process is a Special Model, where the variance tensor V on the n- or $n+1$-dimensional manifold M satisfies $V = \tau^2 I$ i.e., has only one eigenvalue. We have called these models the Simple Models, because of their lack of structure in the prior information. We shall suppose that M satisfies the separation criterion, also in this section. It is not necessary to specify a direction space $\mathcal{L}$ for a Special Model if the variance tensor V is a multiple of the identity. Different direction spaces yield observationally equivalent coefficient processes. There is a chart, such that one of the observationally equivalent coefficient models has a completely linear transition equation w.r.t. this chart if and only if M has zero curvature. In the case that M has non-zero

curvature, there is in general no strong advantage [5] in choosing another chart than the natural one:

$$M \ni q = e^{-r}(x^n - a_1 x^{n-1} + \cdots - a_n) \mapsto (a_1, \cdots, a_n, r) \in V_{nat} \subset \mathbf{R}^{n+1}.$$

Zero-curvature models (with fixed noise levels) allow the interpretation of the Extended Kalman Filter as the Gauss-Newton iteration, given by Fahrmeir/Kaufmann's method in which steps 3 and 5 have been canceled. Non-zero curvature simple models do not permit this interpretation. We shall show this below.

6.4.1 Coefficient identification

The simple model has the following form after the (first) linearization, given in equation (3.34) in chapter three:

$$\begin{aligned} a_{t+1} &= a_t + \tau g(a_t)\lambda_t \\ r_{t+1} &= r_t + \tau\tilde{\lambda}_t \\ z_t &= \phi_t' a_t + e^{-r_t}\eta_t \\ a_0 &\stackrel{d}{=} \mathcal{N}(a_{-1}, Q^2) \quad (Q > 0) \\ r_0 &\stackrel{d}{=} \mathcal{N}(r_{-1}, q^2) \end{aligned} \tag{6.150}$$

where $(\lambda_t)_{t\in\mathbf{Z}}$, $(\tilde{\lambda}_t)_{t\in\mathbf{Z}}$ and $(\eta_t)_{t\in\mathbf{Z}}$ are independent sequences of Gaussian unit variance white noise, and $\phi_t = \begin{pmatrix} y_{t-1} \\ \vdots \\ y_{t-n} \end{pmatrix}$. The square matrix function $g(a_t)$ satisfies $g(a_t)g(a_t)' = H(a_t)^{-1}$. The choice of the factorization gg' of H^{-1} is not important as is clear from what has been said in the first paragraph. This will also be clear from the equations of the Extended Kalman Filter: only the full matrix $H(\hat{a}_t)^{-1}$ will enter these equations. Hence, below we use $g(a_t) = H(a_t)^{-\frac{1}{2}}$.
The criterion function $J_T(a, r)$ for coefficient identification will be minus the logarithm of the posterior density w.r.t. this approximate model (6.150). $J_T(a, r)$ is now not exactly the negative logarithm of the joint density
$p(y_0, \cdots, y_T, a_{0,1}, \cdots, a_{0,n}, r_0, \cdots, a_{T,1} \cdots, a_{T,n}, r_T)$, but only an approximation. Yet, effects of the metric can clearly be pointed out in this criterion. Then, J_T can be expressed

[5] We can indicate some reasons that motivate to take other charts:

1. As Lemma 3.14 in chapter three shows, the error, made in linearizing the transition equation in conformity with one chart, is not the same as with another chart. Hence, one can look for charts for which the error is small, e.g. a normal chart around q_{-1} or around any other point through which the coefficient process passes. One can also elaborate an identification method based on a chain of different normal charts. In such a method the approximation algorithms of section 6.2 can be used.

2. One problem with the method presented below, is that it does not guarantee the estimates to remain in the domain of the natural parametrization. Hence, one can choose a chart $(\tilde{U}_\alpha \subset M, x_\alpha : \tilde{U}_\alpha \to V_\alpha)$ such that $V_\alpha = \mathbf{R}^{n+1}$.

in the natural coordinates:

$$\begin{aligned} J_T &= p_{-1} + \sum_{t=0}^{T-1}(l_t + p_t) + l_T \\ l_t &= r_t + \frac{1}{2}e^{-2r_t}(y_t - \phi_t' a_t)^2 \\ p_t &= \frac{1}{2\tau^2}\langle a_{t+1} - a_t, H(a_t)(a_{t+1} - a_t)\rangle + \frac{c^2}{2\tau^2}|r_{t+1} - r_t|^2 \\ p_{-1} &= \frac{1}{2}||Q^{-1}(a_0 - a_{-1})||^2 + \frac{1}{2}q^{-2}|r_0 - r_1|^2. \end{aligned} \tag{6.151}$$

Its derivative, denoted by $f = (f_0, \cdots, f_T)$ with respect to the coefficients $a_0, \cdots, a_T$, is given by:

$$f_t = \frac{\partial}{\partial a_t} l_t + \frac{\partial}{\partial a_t} p_t + \frac{\partial}{\partial a_t} p_{t-1} \quad (0 \le t \le T-1); \quad f_T = \frac{\partial}{\partial a_T} l_T + \frac{\partial}{\partial a_T} p_{T-1}.$$

This yields:

$$f_0 = e^{-2r_0}(y_t - \phi_0' a_0)\phi_0 - \frac{1}{\tau^2}H(a_0)(a_1 - a_0) -$$

$$-Q^2(a_{-1} - a_0) + \frac{1}{2\tau^2}(a_1 - a_0)' DH(a_0)(a_1 - a_0);$$

$$f_t = e^{-2r_t}(y_t - \phi_t' a_t)\phi_t - \frac{1}{\tau^2}H(a_t)(a_{t+1} - a_t) -$$

$$-\frac{1}{\tau^2}H(a_{t-1}(a_{t-1} - a_t) + \frac{1}{2\tau^2}(a_{t+1} - a_t)' DH(a_t)(a_{t+1} - a_t)$$

if $1 \le t \le T-1$;

$$f_T = e^{-2r_T}(y_T - \phi_T' a_T)\phi_T - \frac{1}{\tau^2}H(a_{T-1})(a_{T-1} - a_T).$$

We only elaborate the situation, where noise levels σ_t^2 are treated as fixed. The Extended Kalman Filter applied to model (6.150) is the same as the Kalman Filter applied to the model:

$$\begin{aligned} a_{t+1} &= a_t + \tau H(\hat{a}_t^p)^{-\frac{1}{2}}\lambda_t \\ z_t &= \phi_t' a_t + \sigma_t \eta_t \\ a_0 &\stackrel{d}{=} \mathcal{N}(a_{-1}, Q^2) \quad (Q > 0), \end{aligned} \tag{6.152}$$

where the point of linearization $\hat{a}^p$ satisfies $\hat{a}_t^p = a_{t|t-1}$, the prediction of a_t by the E.K.F.. Let U be the second derivative of the log-posterior density $p(a_0, \cdots, a_T | z_0, \cdots, z_T)$ of the linear system (6.152). The matrix U is block-tridiagonal. The non-zero block elements of U are:

$$\begin{aligned} U_{00} &= \frac{\phi_0 \phi_0'}{\sigma_0^2} + \frac{1}{\tau^2}H(a_0) + Q^2; \\ U_{tt} &= \frac{\phi_t \phi_t'}{\sigma_t^2} + \frac{1}{\tau^2}H(a_t) + \frac{1}{\tau^2}H(a_{t-1}) \quad (1 \le t \le T-1); \end{aligned}$$

$$U_{TT} = \frac{\phi_T \phi_T'}{\sigma_T^2} + \frac{1}{\tau^2} H(a_{T-1});$$

$$U_{t-1\ t} = U_{t\ t-1} = -\frac{1}{\tau^2} H(a_{t-1}) \quad (1 \leq t \leq T). \tag{6.153}$$

The second derivative of the criterion function J_T with respect to a_s, a_t, however, is different from U. In order to obtain equality, we have to drop those terms that contain derivatives of $H(a_t)$, terms of the form $\frac{1}{\tau^2} DH(a_t)(a_{t+1} - a_t)$ and $\frac{1}{2\tau^2}(a_{t+1} - a_t)' D^2 H(a_t)(a_{t+1} - a_t)$. The exclusion of these terms makes that an iteration step that uses gain matrix U, can't be viewed as a Gauss-Newton method, at most as a very modified Newton method. Yet, the matrix U is positive definite, because the *controllability gramians* $\mathcal{K}(t,t) = \tau^2 H(\hat{a}_t^p)^{-1}$ are all positive definite for $0 \leq t \leq T$, which implies that the prior variance of $(a_0, \cdots, a_T)$ in model (6.152) is positive definite. Accordingly, Fahrmeir/Kaufmann's method is feasible using matrix U, and the Extended Kalman Filter is just this method where steps 3 and 5 are not performed. In applying the Extended Kalman Filter or Fahrmeir/Kaufmann's method, there may be one problem: the estimates $\hat{a}_0^{t_1}, \cdots, \hat{a}_{t_1}^{t_1}$ can run out of their domain $\frac{1}{\sigma_0} AR_n \cap M, \cdots, \frac{1}{\sigma_{t1}} AR_n \cap M$ respectively. The main problem in executing the method after the estimate has run out of its domain is that the matrix $H(a)$ is not defined anymore. The analytic expressions for $H(a)$ may still be defined, but can yield, for instance, a negative definite matrix. There are several ways to resolve this problem. One way is to define a Riemannian structure on the complement of AR_n in $\mathbb{R}^n$ too.

6.4.2 Identification of hyperparameters

Next, we shall discuss methods to determine the hyperparameters of the simple model with constant noise levels σ^2. These are in fact :

- the hyperparameters, determining the initial distribution $a_0 \stackrel{d}{=} \mathcal{N}(a_{-1}, Q^2)$;
- the measurement noise level σ^2;
- the noise-variance-ratio $NVR = \frac{\tau^2}{\sigma^2}$;
- the order n.

System (6.152) will be uniformly completely observable if, for some $N \in \mathbb{N}$, the observability gramian

$$\mathcal{J}(t+N, t+1) = \sum_{k=1}^{N} \frac{\phi_{t+k} \phi_{t+k}'}{\sigma^2} \tag{6.154}$$

is uniformly bounded away from zero and infinity for all $t \in \mathbb{N}$. The system will be uniformly completely controllable if, for some $N \in \mathbb{N}$, the controllability gramian

$$\mathcal{K}(t+N, t+1) = \tau^2 \sum_{k=1}^{N} H(\hat{a}_{t+k}^p)^{-1} \tag{6.155}$$

is uniformly bounded away from zero and infinity for all $t \in \mathbb{N}$. We refer to Jazwinski (1970), chapter 7, for these concepts. The first concept depends completely on the data. The second depends on the chosen metric and via the trajectory of the estimated coefficients on the data. If the system is uniformly completely observable and uniformly completely controllable, then the Kalman Filter applied to the system (6.152) will be asymptotically stable. This implies that estimates $a_{0|-1}$ of the initial value a_0 and its error variance $P_{0|-1}$ are gradually forgotten for a larger number of data $(y_t)_{0 \leq t \leq T}$, see theorems 7.4 and 7.5 in Jazwinski (1970). Hence, if system (6.152) is uniformly completely observable and uniformly completely controllable, then the determination of the initial values is a problem of minor importance for large data sets.
The remaining hyperparameters σ^2, NVR and the order n can be estimated following the same procedure as in case of the geodesic model. Also here, we work with an *approximate* likelihood, obtained by Prediction Error Decomposition by the Kalman Filter applied to the linearized model (6.152). We can form the AIC criterion by adding $2n$ to the $-2\log$-likelihood, to estimate the order of the autoregression. In the same way as explained for the geodesic model, the noise-level σ^2 can be concentrated out of the AIC criterion. This AIC criterion can then be minimized with respect to the NVR and the order n by discrete parameter search. This method is quite standard for linear time-varying dynamical models; for instance, see Kitagawa and Gersch (1985).

6.4.3 Experiments

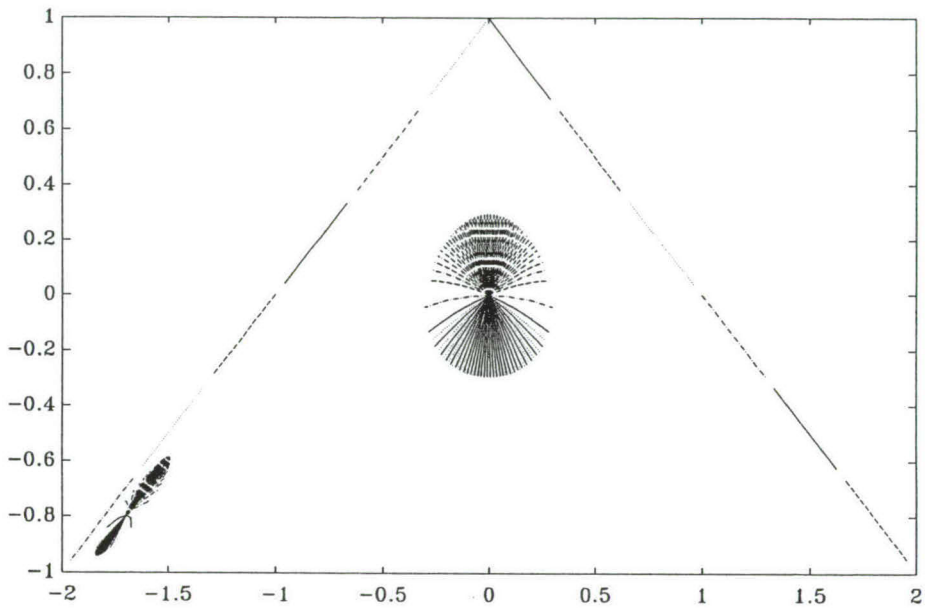
The simple model and the geodesic model have been subjected to a large number of simulation experiments. For a large part, these experiments were performed at the university of Lancaster. The outline of the experiments was that a data series was generated, according to some general model of the complexity-types 2 and 3 as described in chapter three, section 3.9. Complexity-type 1 corresponds to the simple models, complexity-type 2 to the geodesic models and complexity-type 3 to models with two non-zero eigenvalues of the variance tensor field V. The manifold was nearly always $\frac{1}{\sigma} AR_2 \setminus \{q_{-1}\}$, and the direction space $\mathcal{L}$ contained the tangent field F of the geodesics through the point q_1. The asymptotic Fisher metric was used, sometimes also the Euclidean. Note that it is possible to generate these series, according to the asymptotic Fisher metric exactly, because we have formulas at our disposal of all geodesics in AR_2. We report on the experiments, where the asymptotic Fisher metric was used to generate the series.
Every data series then was subjected to identification using:

1. a simple model based on the Euclidean metric,

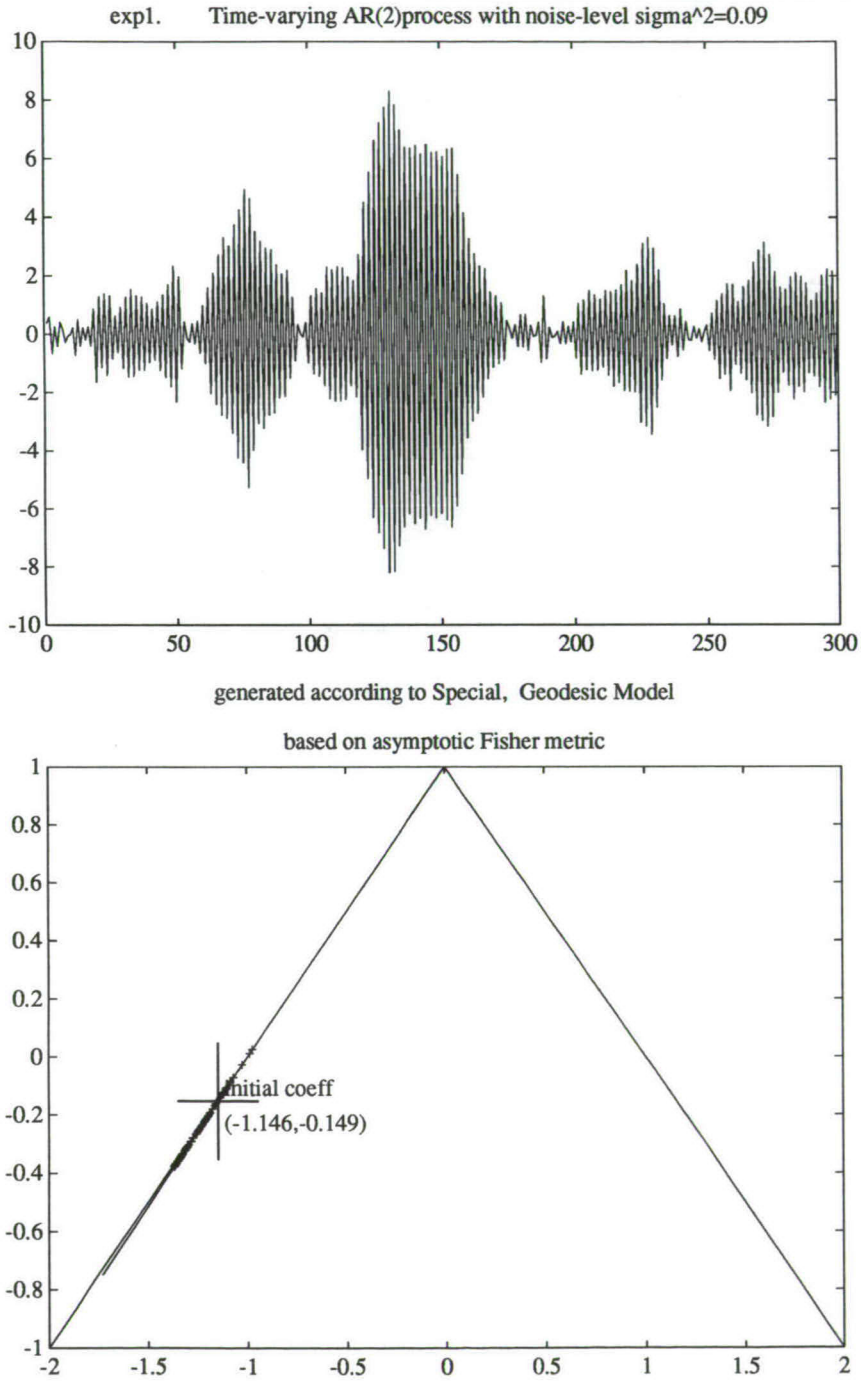
2. a simple model based on the asymptotic Fisher metric and

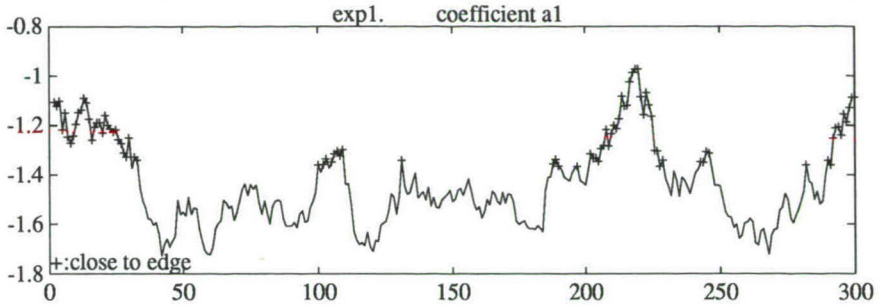
3. a geodesic model (geodesics through the point q_{-1}) based on either the Euclidean or the asymptotic Fisher metric.

It was assumed that the pre-initial point q_{-1} was given. In the identification method, not the formulas of the geodesics were used, but the approximation method of section 6.2. We restricted ourselves to application of the Extended Kalman Filter without smoothing, and applied the methods, given in the preceding sections, to determine the hyperparameters. We report on the results of some of these experiments. In the figures, concerning the identification by a geodesic model, we also depicted the development of the length of the estimated tangent vector field F and the development of its inner product (in the used metric) with the estimate of the vector field $Z = \frac{\partial}{\partial \zeta}$, i.e. the second parameter direction. According to section 6.2, the length $||F||$ should be constantly 1, and the inner product $\langle F, Z \rangle$ should be constantly zero. Hence, looking at the development of these quantities is a diagnostic check, whether the approximation method and the whole model worked all right. The other figures give the picture of the data and the generated curve of the coefficients in the AR(2) stability region in comparison with the curves of estimated coefficients, according to the three different models.
It was found that in general the ML estimate of the NVR, according to the asymptotic Fisher metric-model, was larger than the one, according to the Euclidean model. The data cause the latter filter to be more cautious in view of the times t that the shape of the Euclidean metric is very different from the shape of the asymptotic Fisher metric. Loss of track occurred in the Euclidean filter, when the generated coefficients moved close to the upper edges of the triangle. Here the shape of the quadratic form, defined by the asymptotic Fisher metric, is very different from the quadratic form, defined by the Euclidean metric. This is illustrated in the figure on the next page.
Results overall were reasonable for all three models. In 50% there was no big difference in performance of the three methods; in the other experiments clearly the second was somewhat better than the first and the third somewhat better than the second.

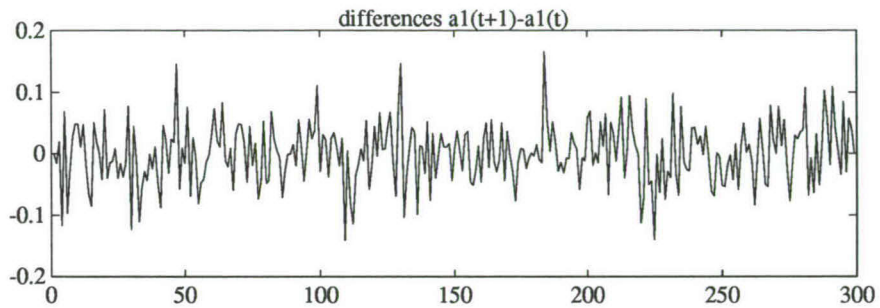


balls of equal size in the asymptotic Fisher metric on AR_2

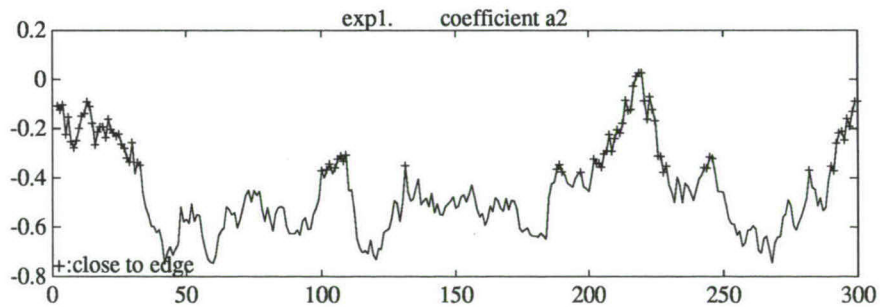

expl. Time-varying AR(2)process with noise-level sigma^2=0.09
10
8
6
4
2
0
-2
-4
-6
-8
-10
0
50
100
150
200
250
300
generated according to Special, Geodesic Model
based on asymptotic Fisher metric
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
Initial coeff
(-1.146,-0.149)
---:curve of coefficients (+++:points of curve close to edge)



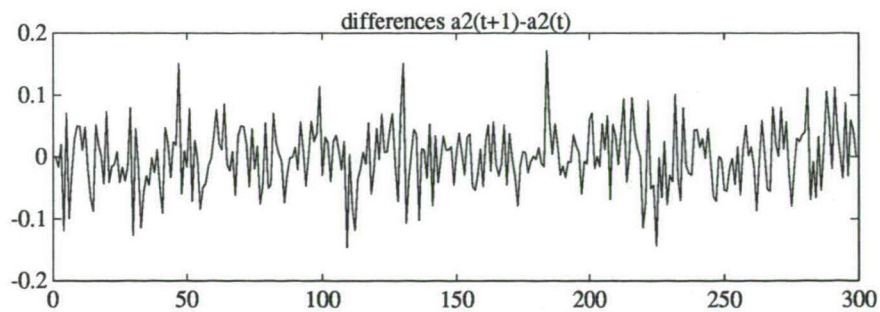
Special, Geodesic Model (asymptotic Fisher metric)



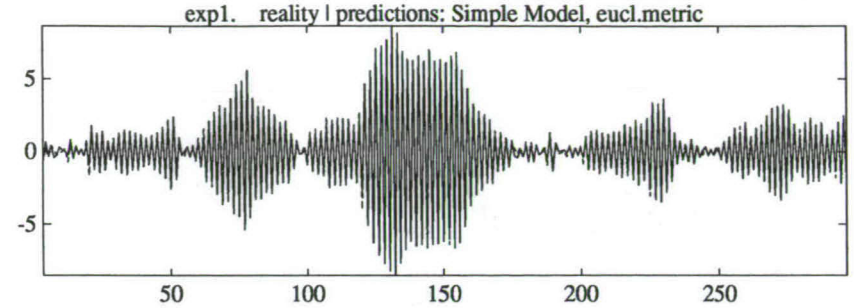
transition hyperparameter in jumps is tau=0.06



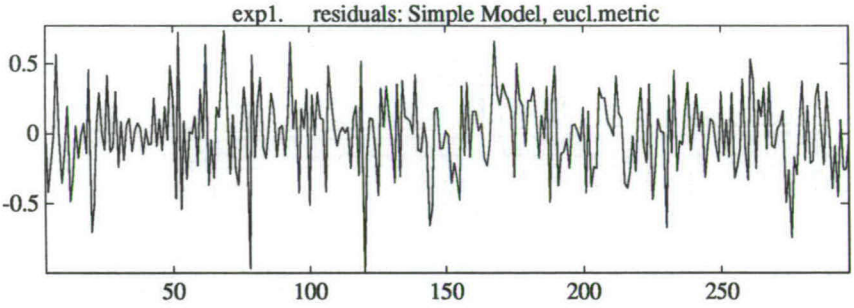
(300 jumps in coeff. on 300 observ.)



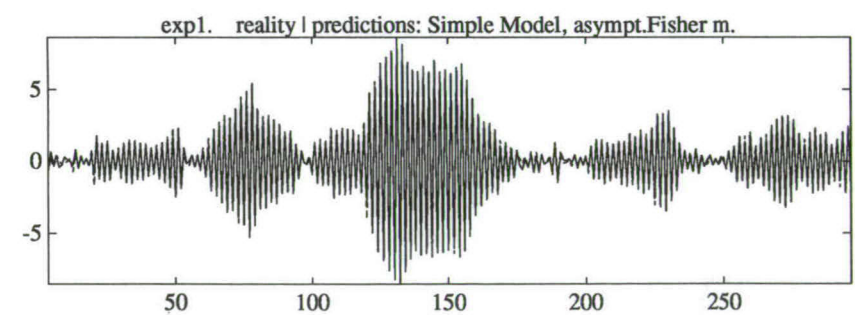
NVR=tau^2/sigma^2=0.04 (Geodesic Model!)



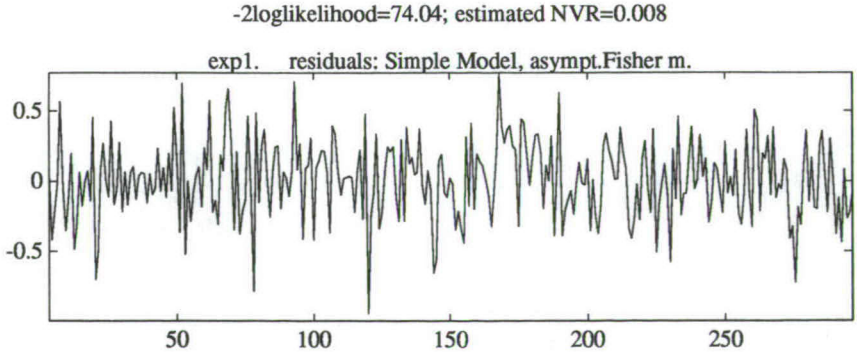
-2loglikelihood=83.97; estimated NVR=0.001



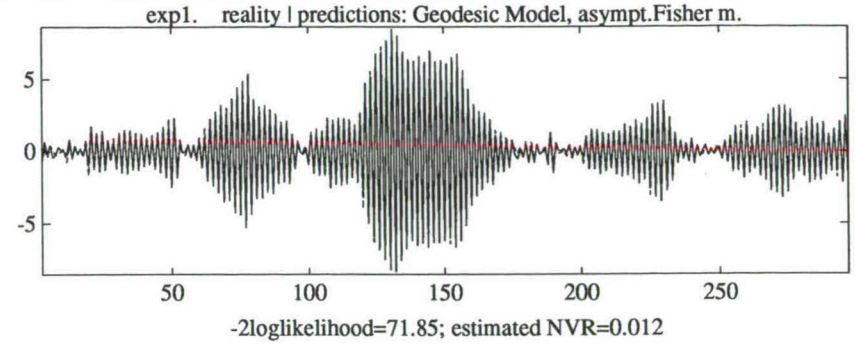
mean=0.005; stand.dev.=0.28



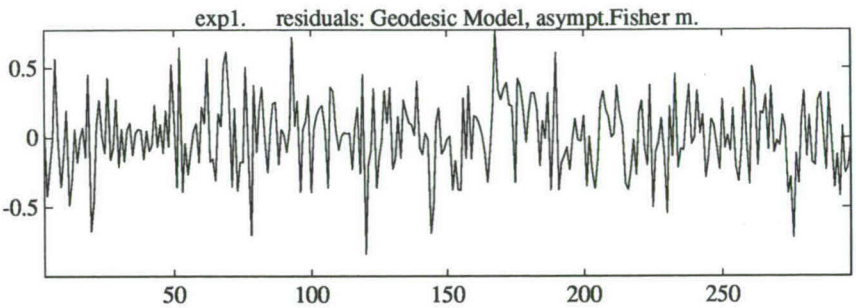
-2loglikelihood=74.04; estimated NVR=0.008



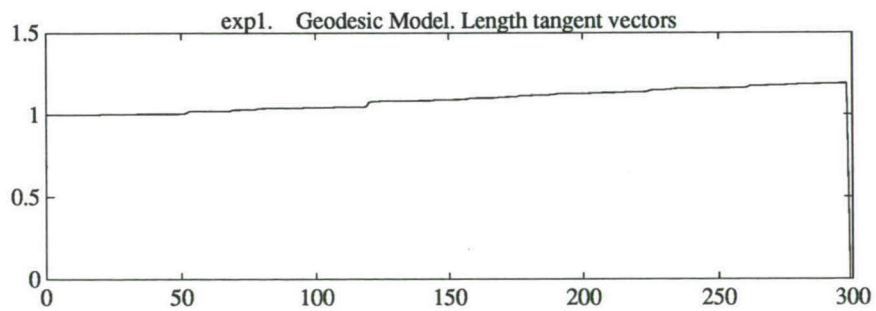
mean=0.008; stand.dev.=0.274

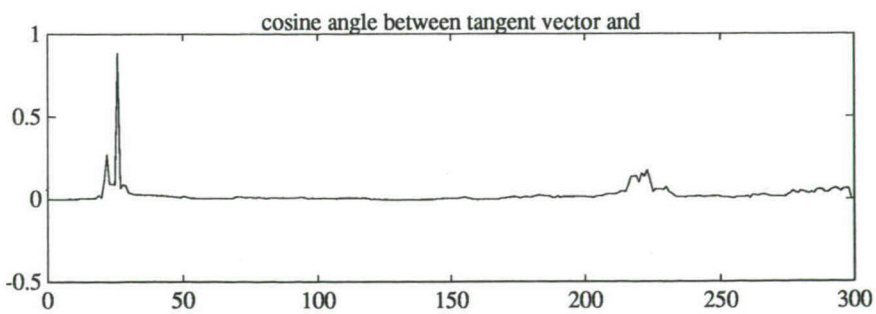


-2loglikelihood=71.85; estimated NVR=0.012



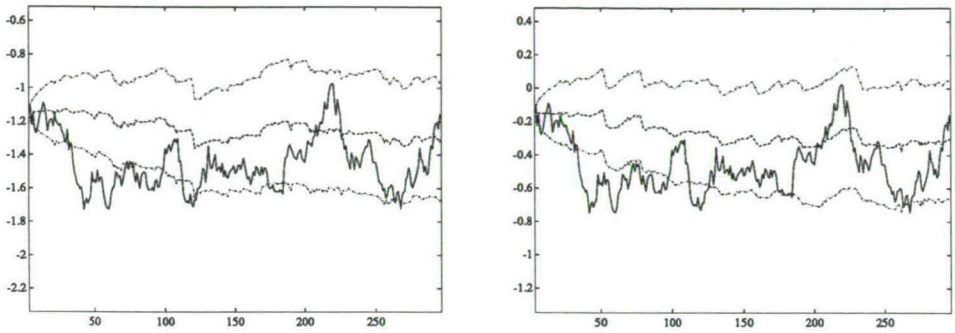
mean=0.016; stand.dev.=0.286



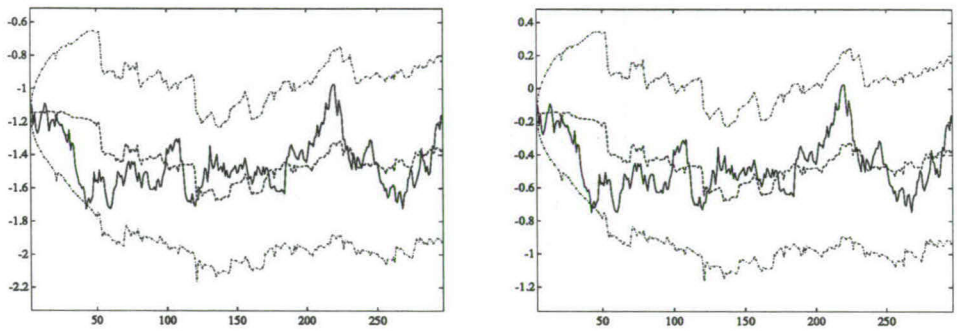


second parameter direction

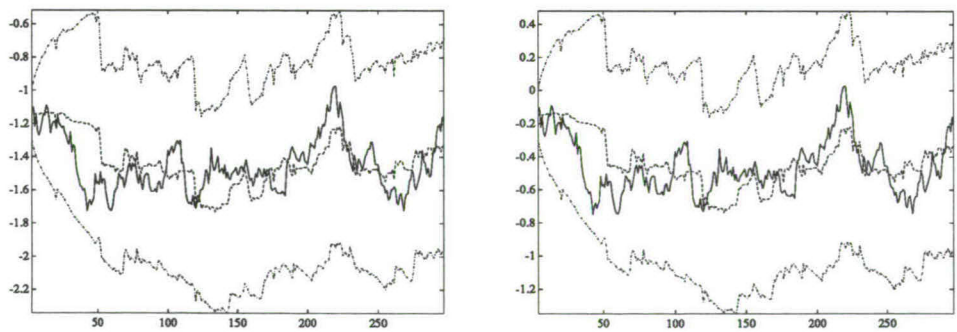
coefficient a_1 coefficient a_2



dashed: euclidean metric (Simple Model)



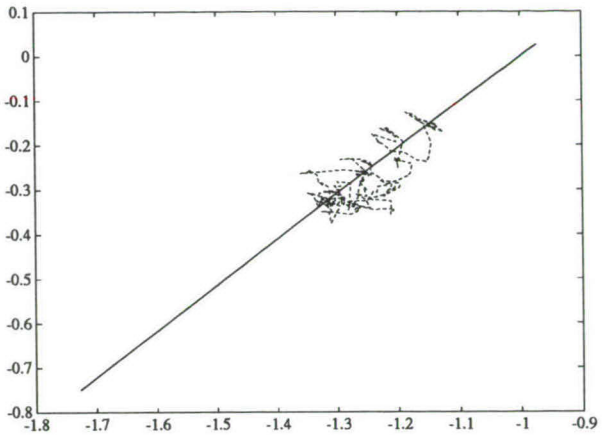
dashed: asymptotic Fisher metric (Simple Model)



dashed: asymptotic Fisher metric (Geodesic Model)

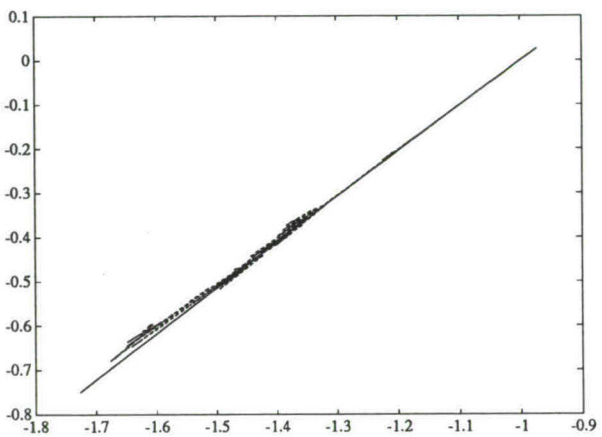
solid: generated coefficients dashdot: plus/min root estimated "error variance"

curve of coefficients a_1 and a_2

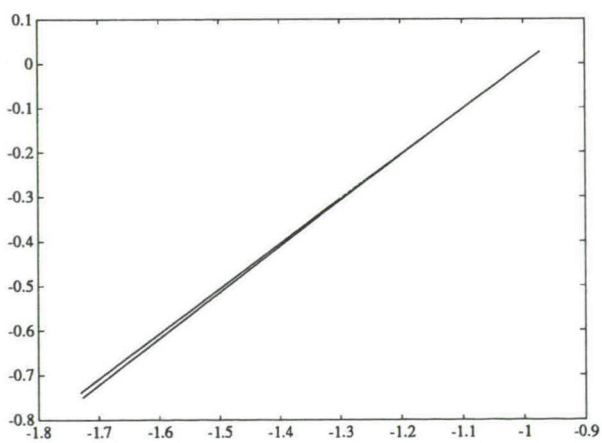


solid: curve of generated coefficients

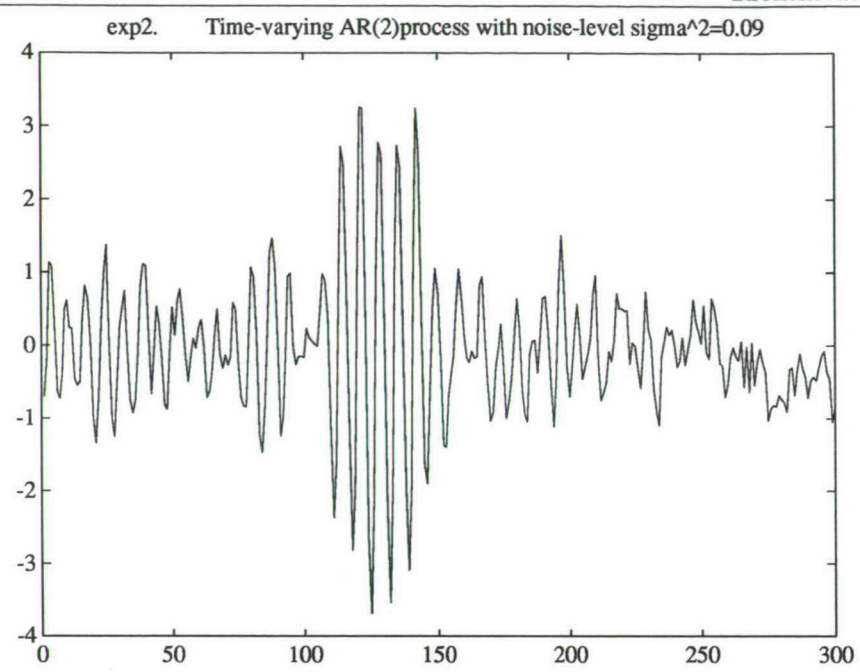
dashed: euclidean metric (Simple Model)



dashed: asymptotic Fisher metric (Simple Model)

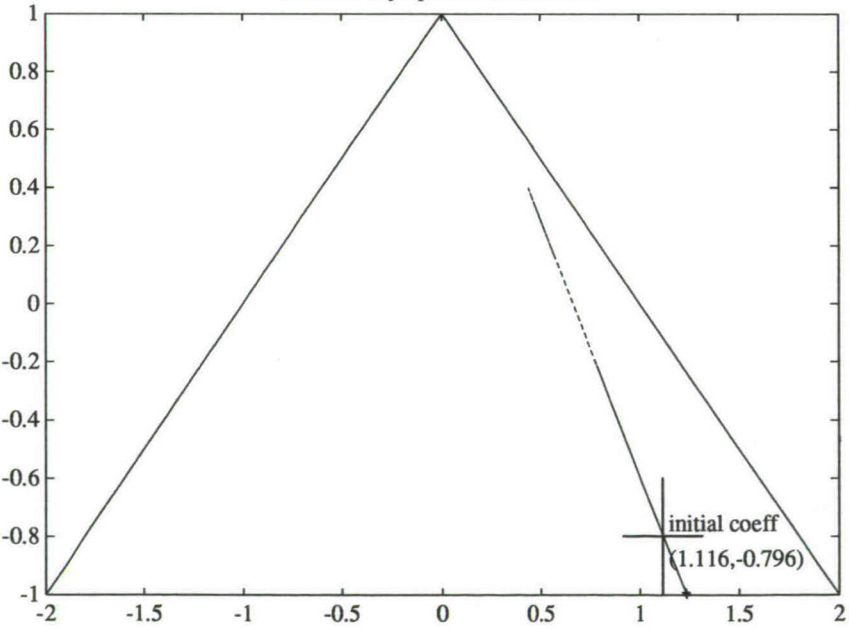


dashed: asymptotic Fisher metric (Geodesic Model)

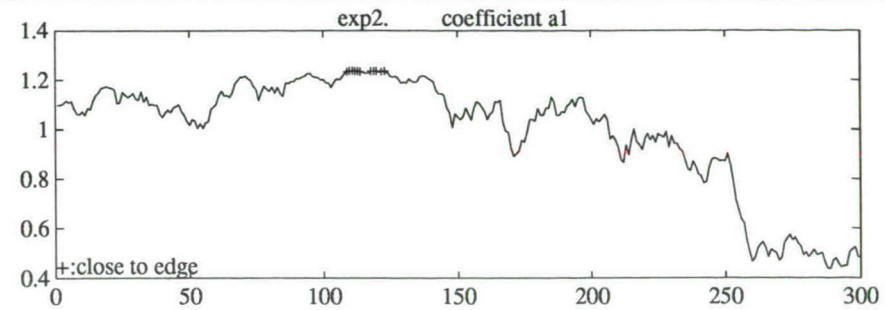


generated according to Special, Geodesic Model

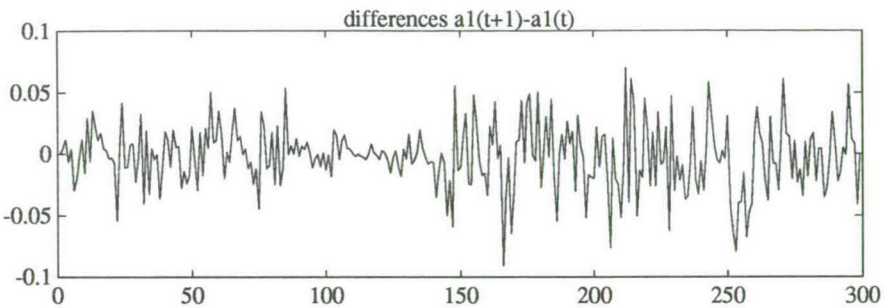
based on asymptotic Fisher metric



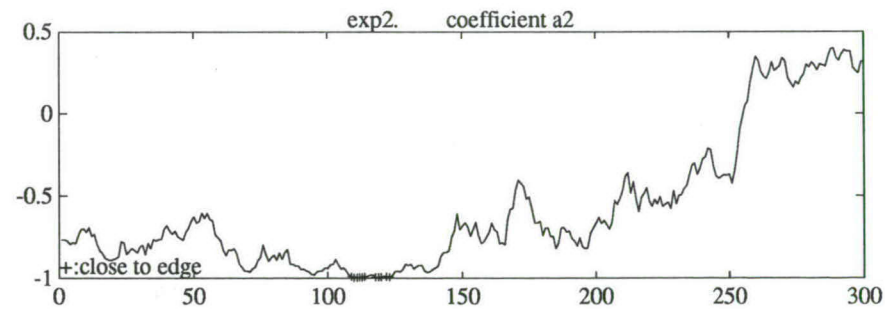
---:curve of coefficients (+++:points of curve close to edge)



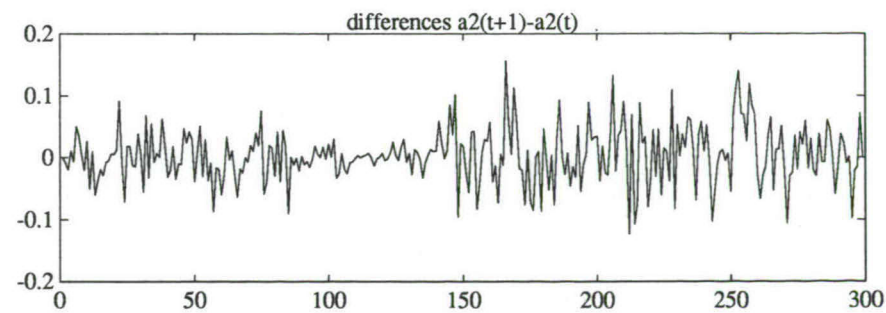
Special, Geodesic Model (asymptotic Fisher metric)



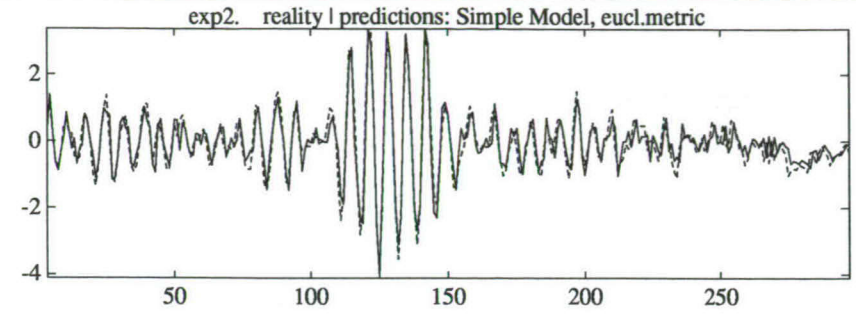
transition hyperparameter in jumps is tau=0.06



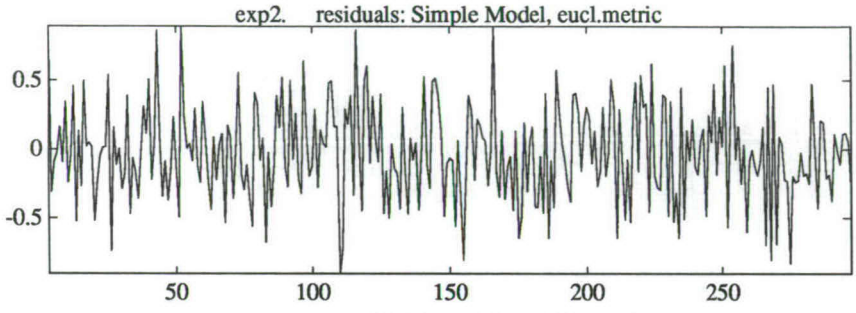
(300 jumps in coeff. on 300 observ.)



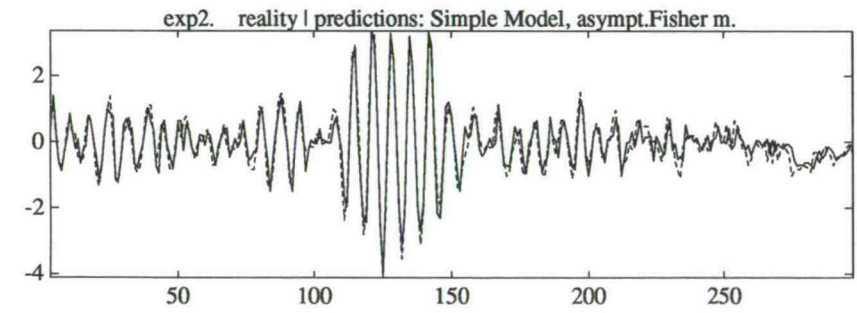
NVR=tau^2/sigma^2=0.04 (Geodesic Model!)



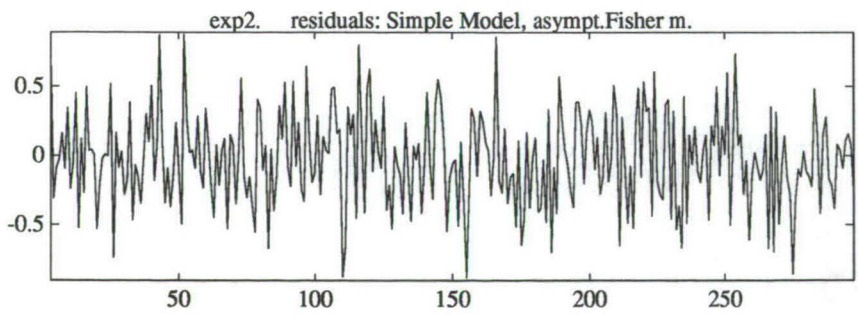
-2loglikelihood=213.9; estimated NVR=0.008



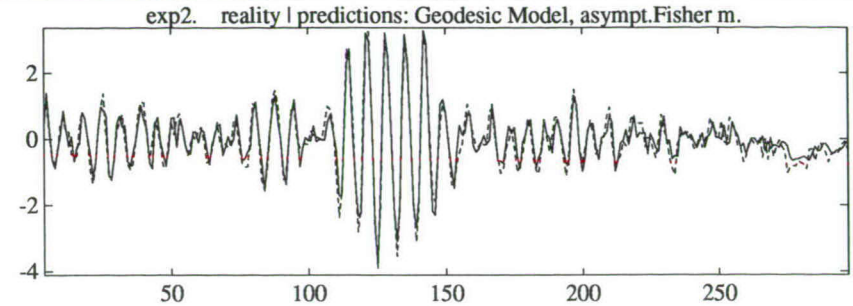
mean=-0.026; stand.dev.=0.344



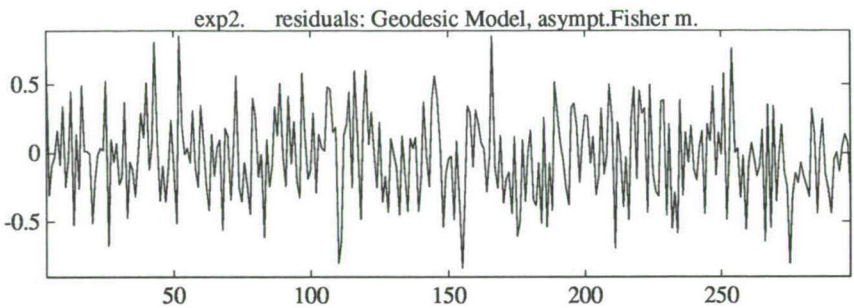
-2loglikelihood=204.2; estimated NVR=0.03



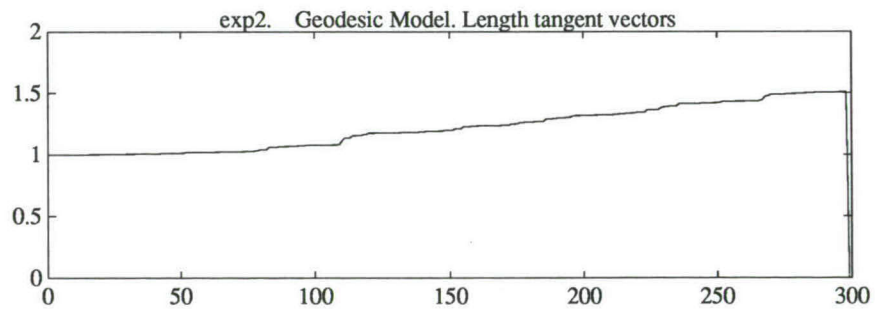
mean=-0.026; stand.dev.=0.338

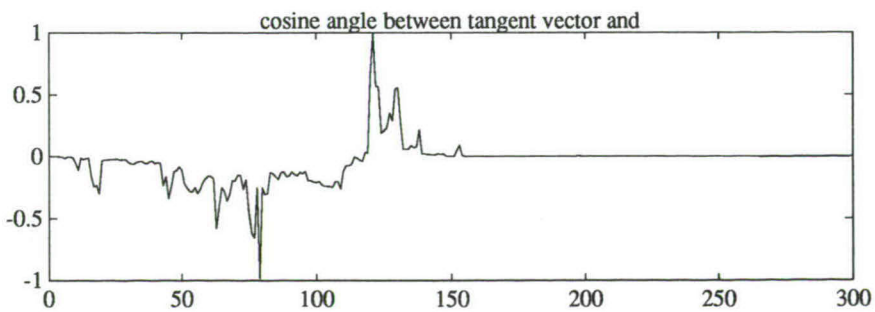


-2loglikelihood=193.6; estimated NVR=0.015



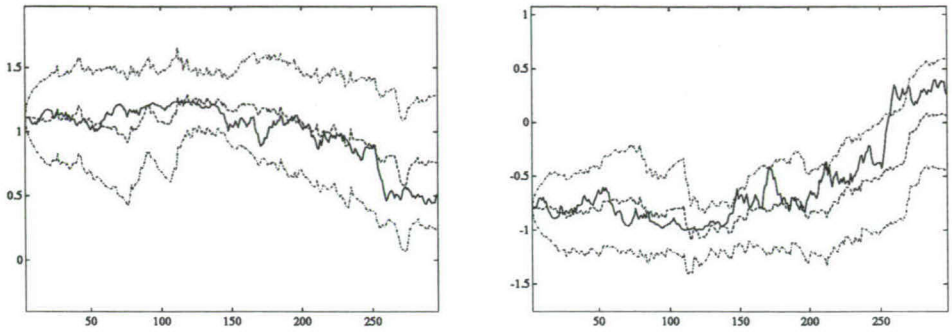
mean=-0.03; stand.dev.=0.322



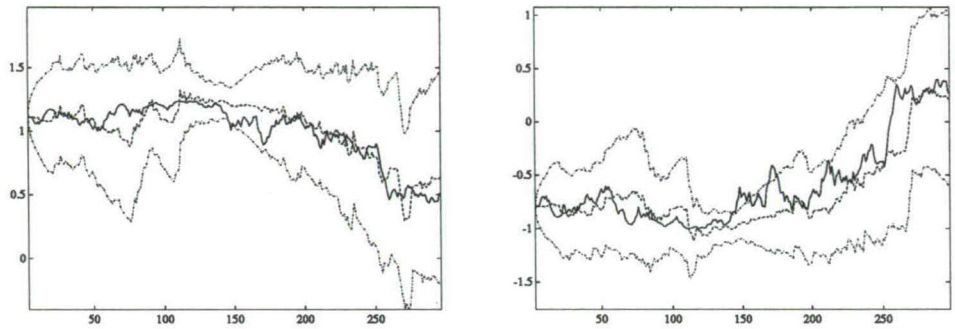


second parameter direction

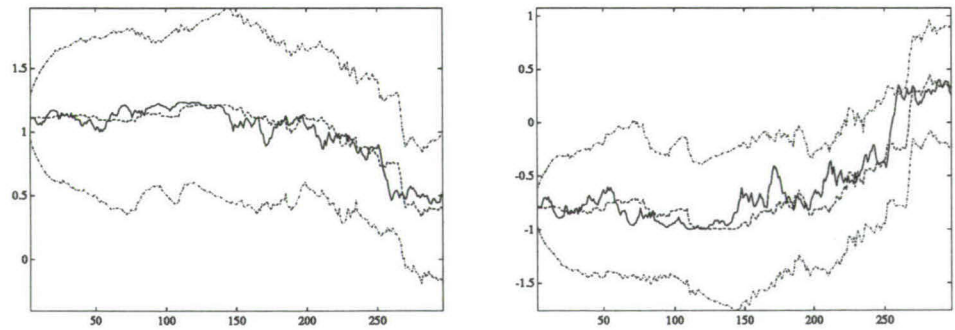
coefficient a_1 coefficient a_2



dashed: euclidean metric (Simple Model)



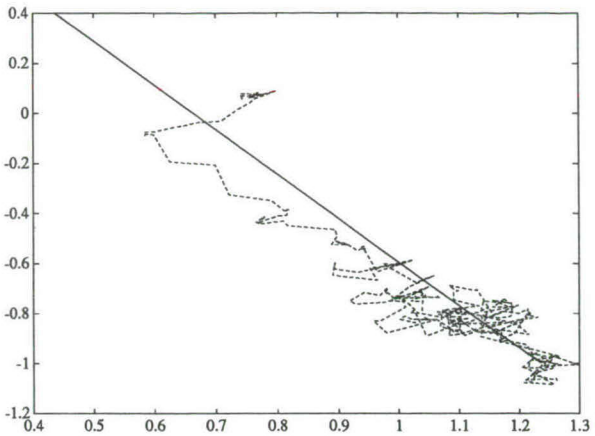
dashed: asymptotic Fisher metric (Simple Model)



dashed: asymptotic Fisher metric (Geodesic Model)

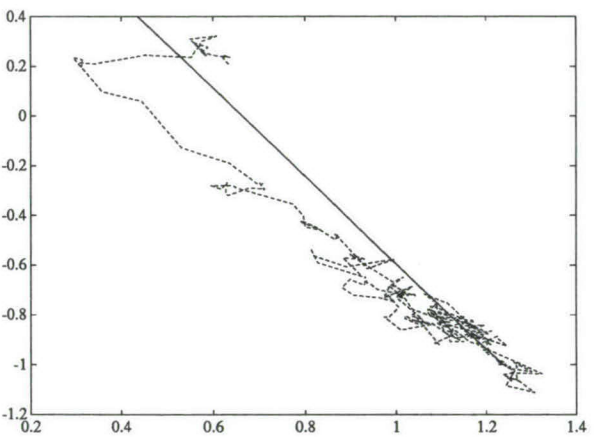
solid: generated coefficients dashdot: plus/min root estimated "error variance"

curve of coefficients a_1 and a_2

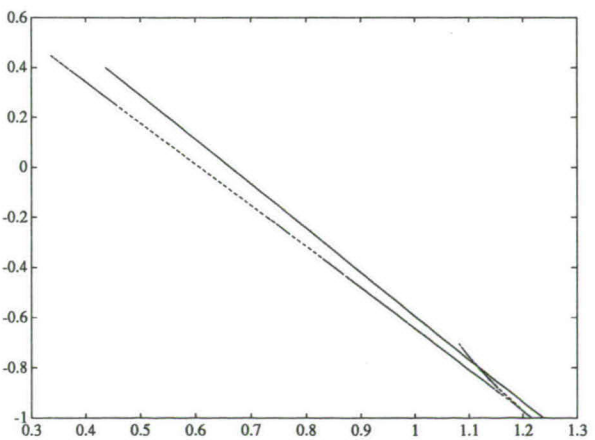


solid: curve of generated coefficients

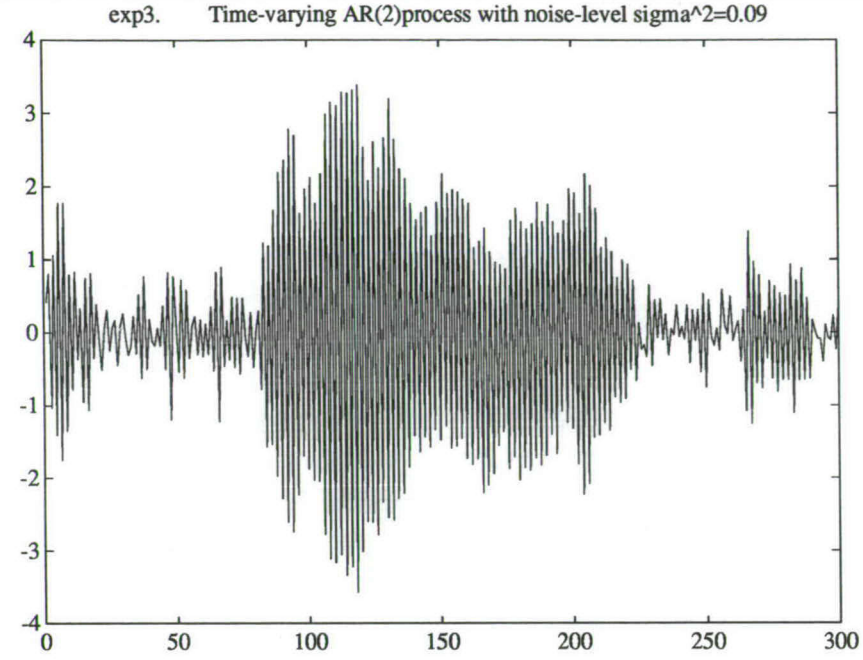
dashed: euclidean metric (Simple Model)



dashed: asymptotic Fisher metric (Simple Model)

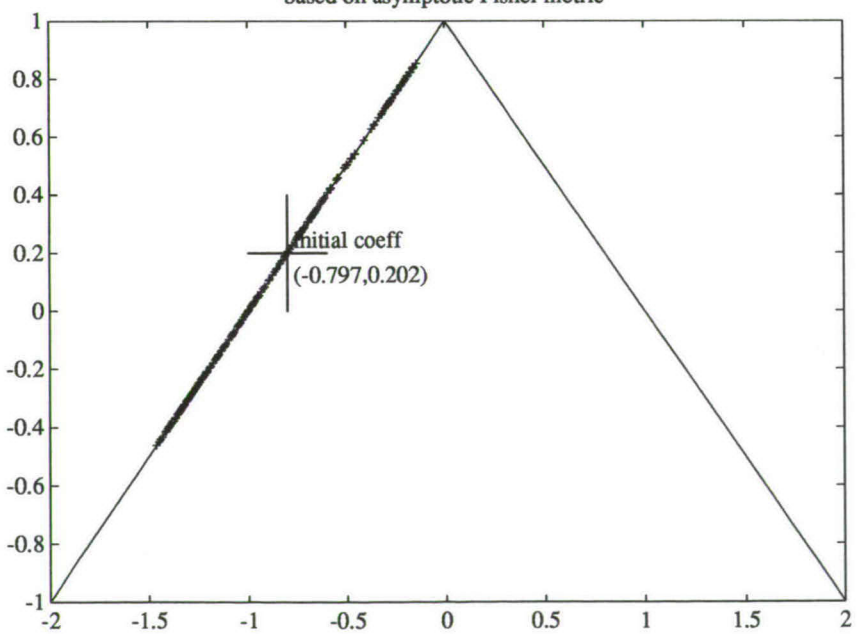


dashed: asymptotic Fisher metric (Geodesic Model)

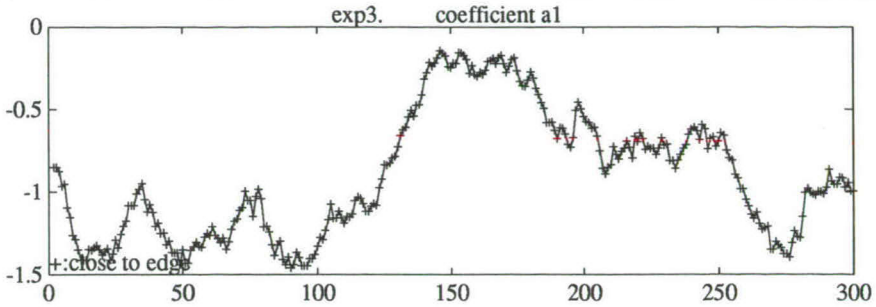


generated according to Special, Geodesic Model

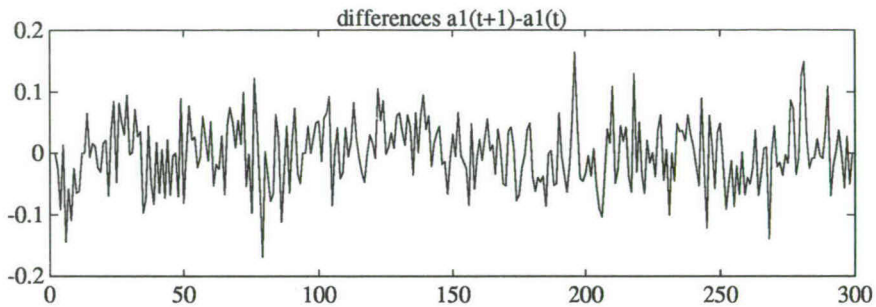
based on asymptotic Fisher metric



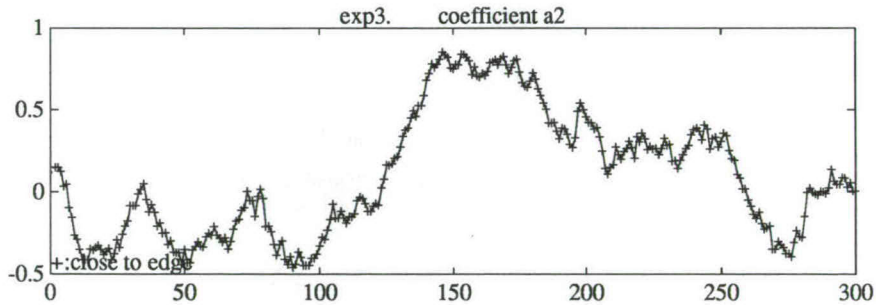
---:curve of coefficients (+++:points of curve close to edge)



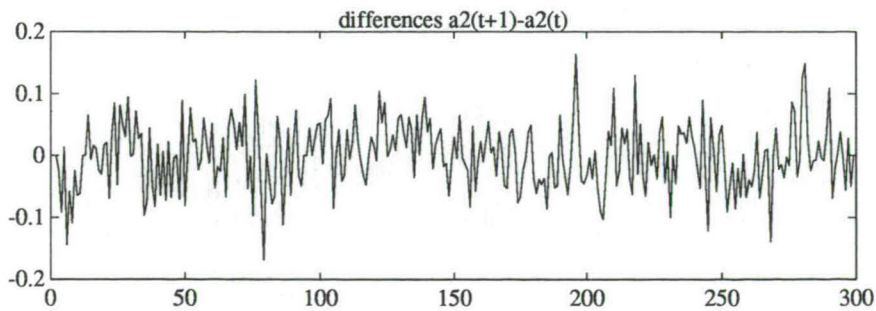
Special, Geodesic Model (asymptotic Fisher metric)



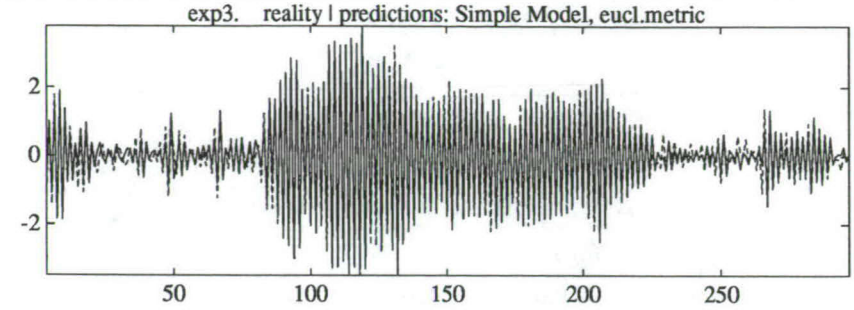
transition hyperparameter in jumps is tau=0.06



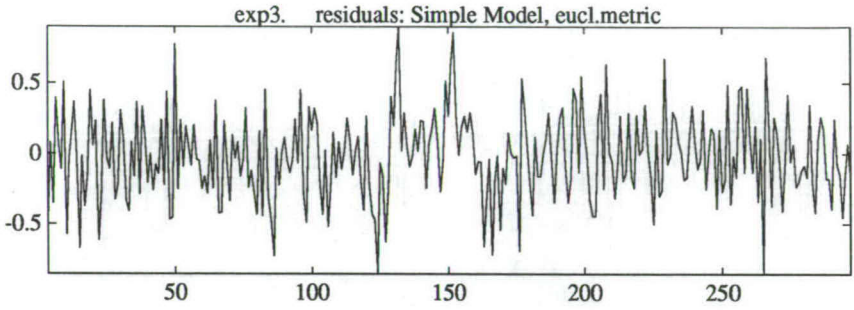
(300 jumps in coeff. on 300 observ.)



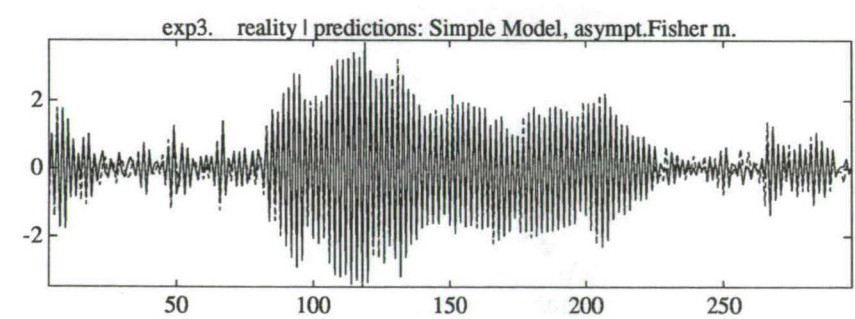
NVR=tau^2/sigma^2=0.04 (Geodesic Model!)



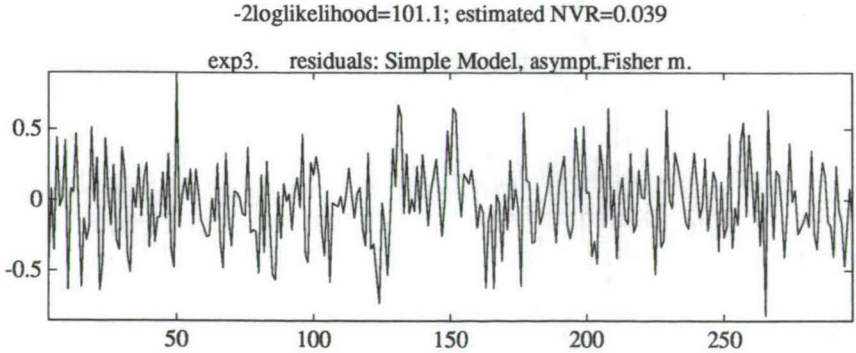
-2loglikelihood=127.2; estimated NVR=0.023



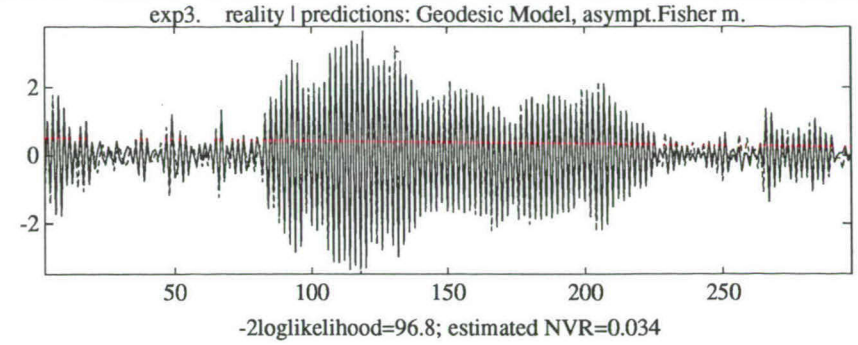
mean=-0.016; stand.dev.=0.301



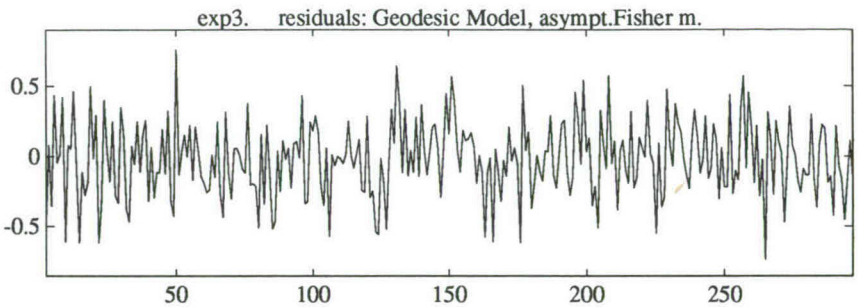
-2loglikelihood=101.1; estimated NVR=0.039



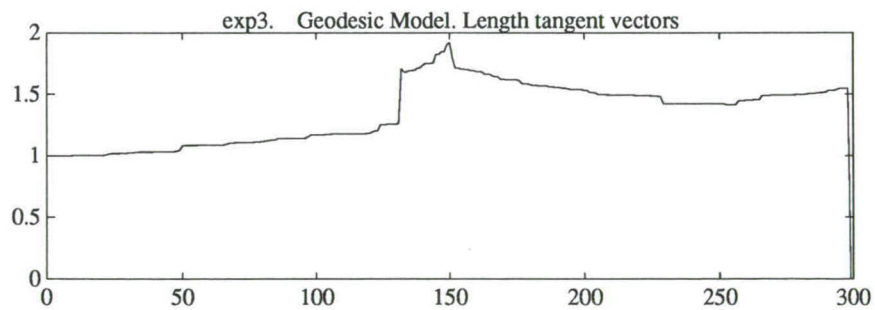
mean=-0.017; stand.dev.=0.286

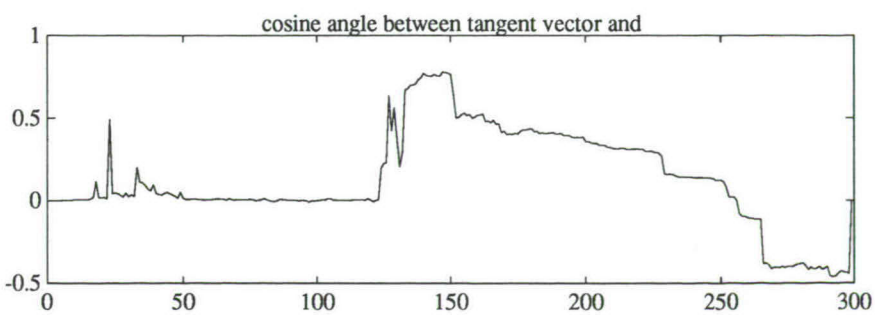


-2loglikelihood=96.8; estimated NVR=0.034

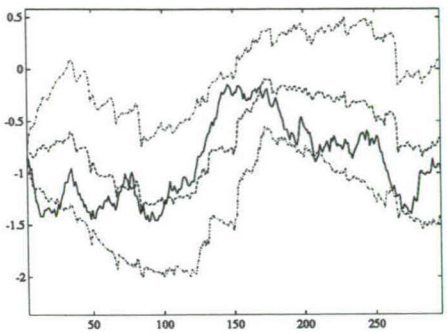


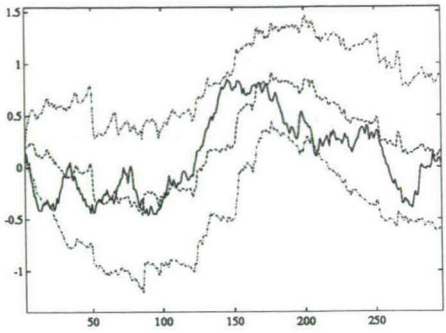
mean=-0.018; stand.dev.=0.266



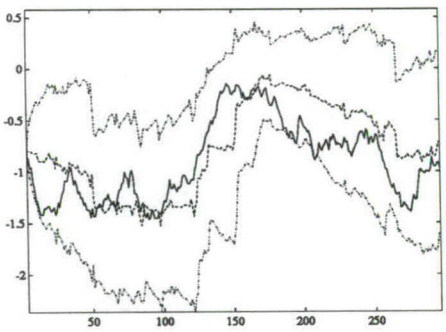


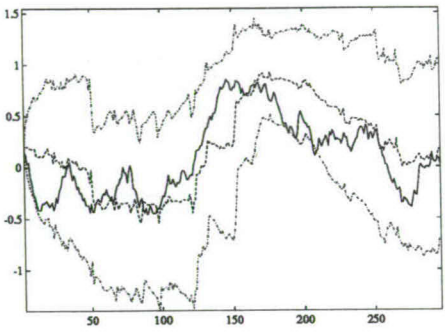
coefficient a_1 coefficient a_2



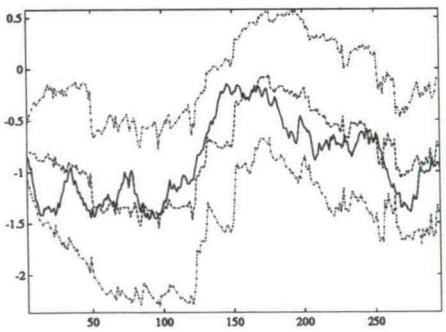


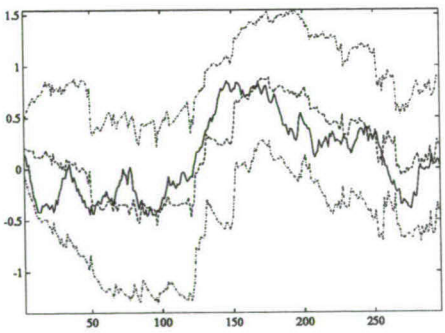
dashed: euclidean metric (Simple Model)





dashed: asymptotic Fisher metric (Simple Model)

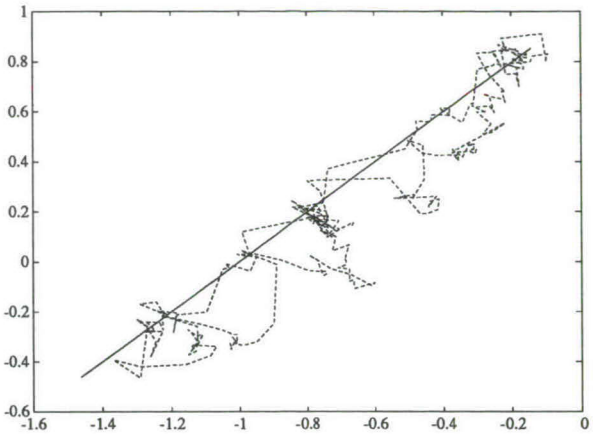




dashed: asymptotic Fisher metric (Geodesic Model)

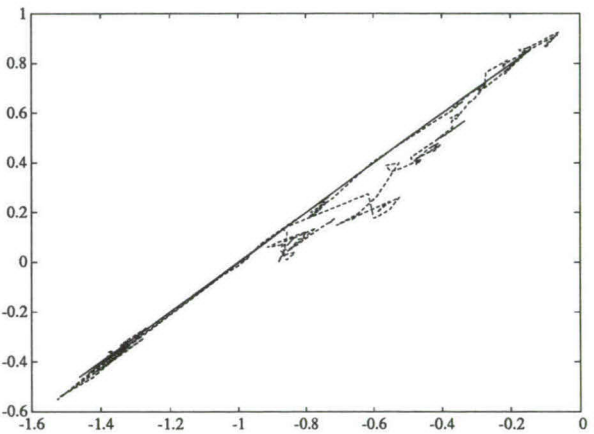
solid: generated coefficients dashdot: plus/min root estimated "error variance"

curve of coefficients a_1 and a_2

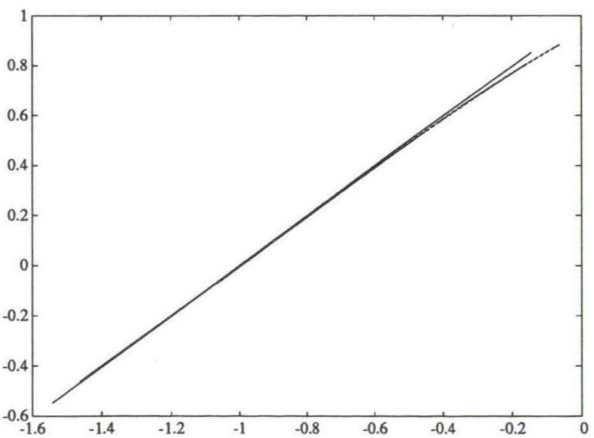


solid: curve of generated coefficients

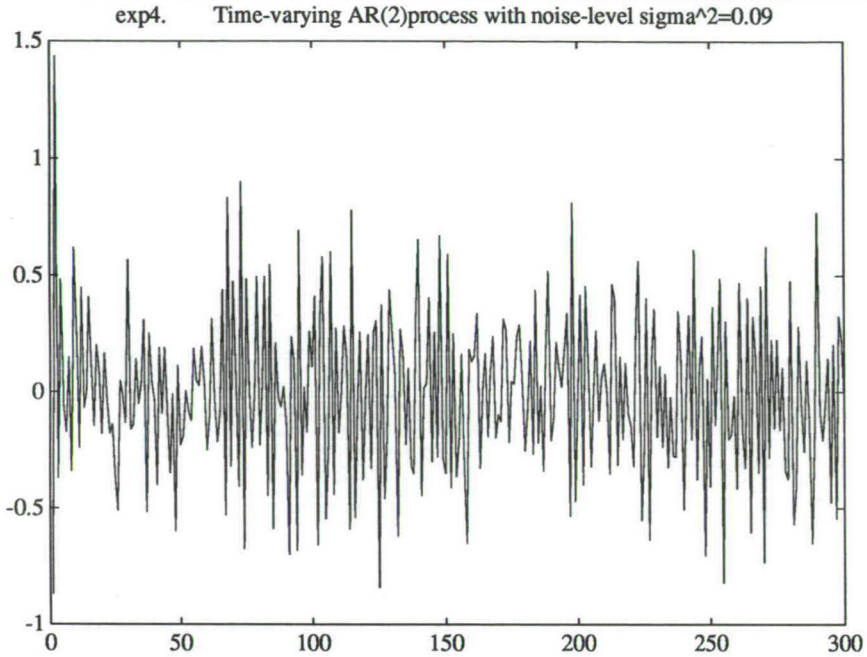
dashed: euclidean metric (Simple Model)



dashed: asymptotic Fisher metric (Simple Model)

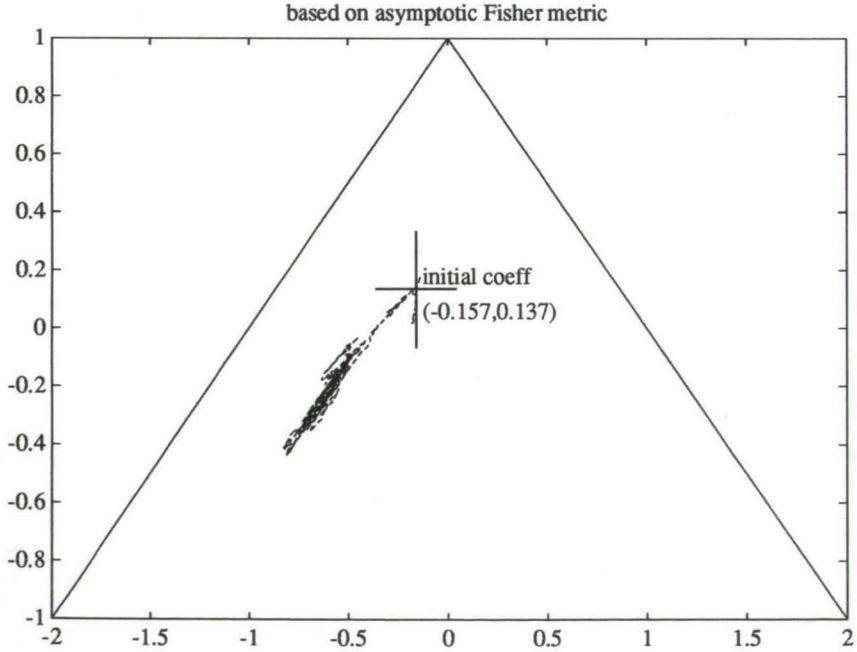


dashed: asymptotic Fisher metric (Geodesic Model)

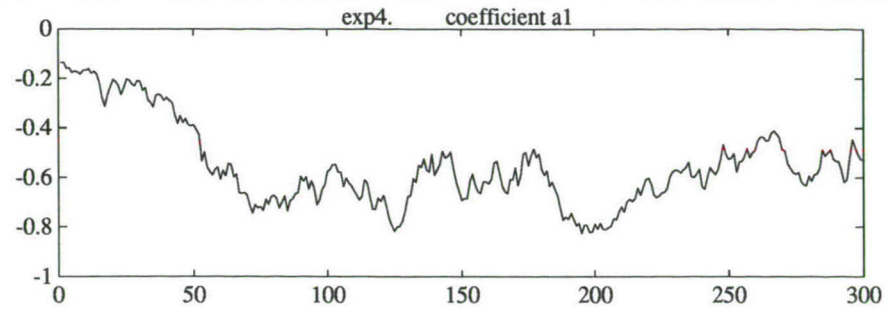


generated according to Special, Model of Complexity type 3

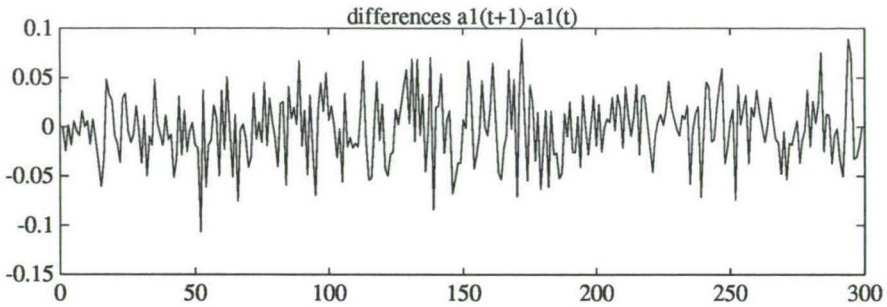
based on asymptotic Fisher metric



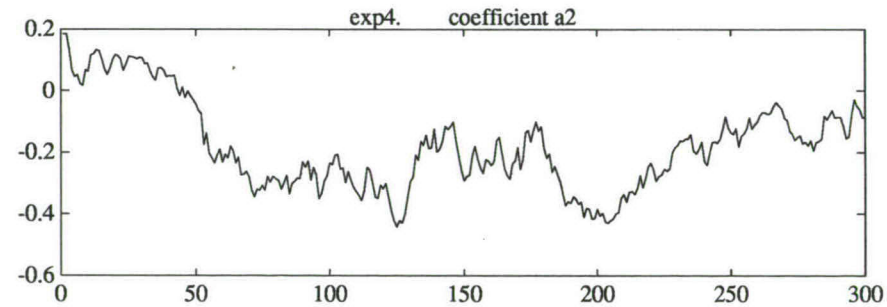
curve of coefficients



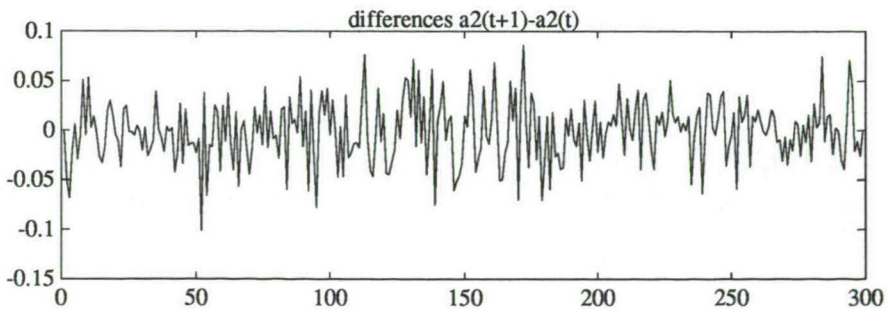
Special, Model of Complexity type 3 (asymptotic Fisher metric)



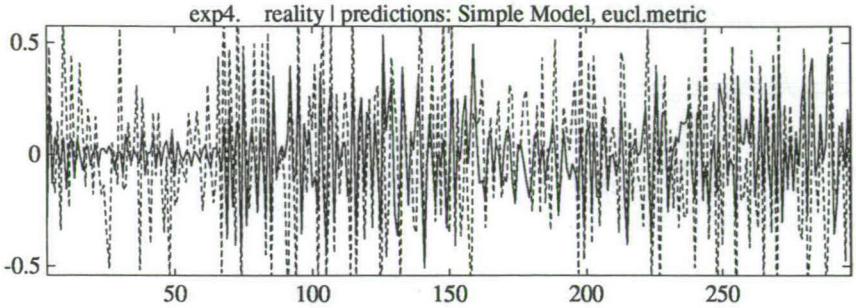
transition hyperparameters in jumps are tau1=0.008; tau2=0.04



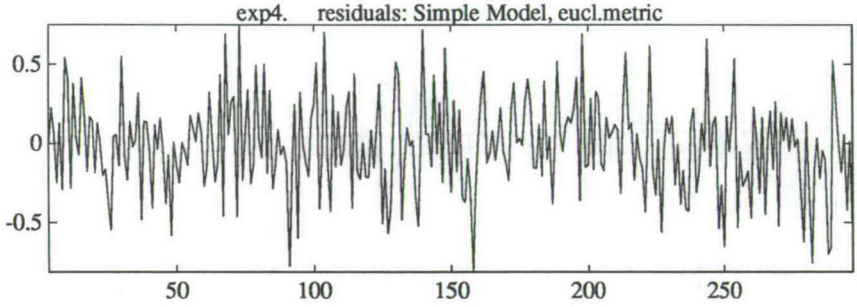
(300 jumps in coeff. on 300 observ.)



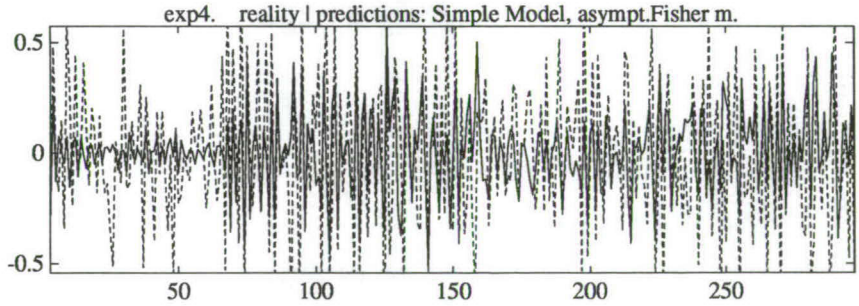
largest NVR=tau2^2/sigma^2=0.017 (Model of Complexity type 3!)



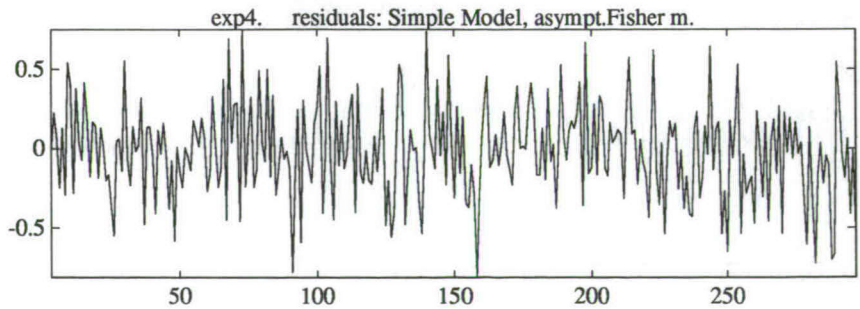
-2loglikelihood=121.5; estimated NVR=0.01



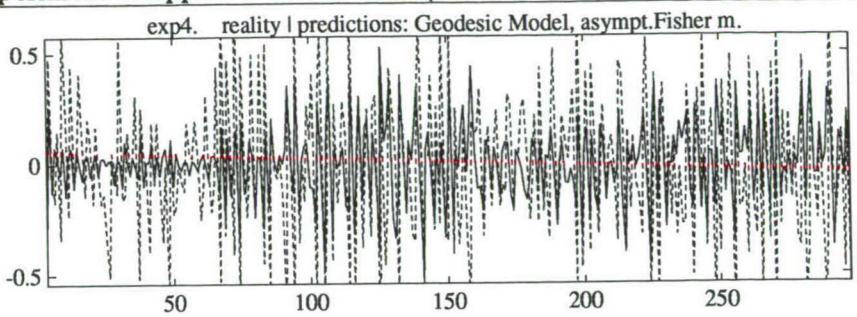
mean=-0.015; stand.dev.=0.295



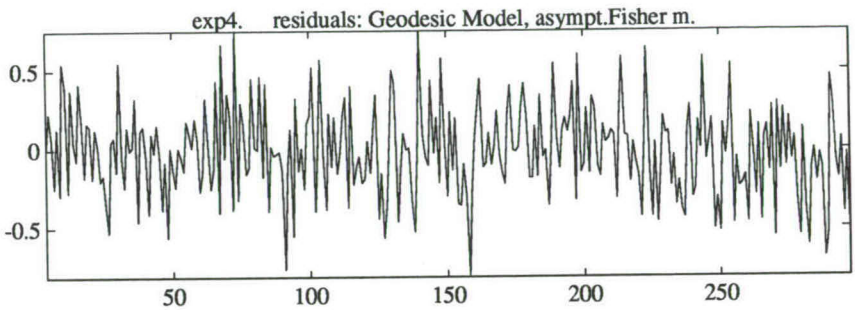
-2loglikelihood=120.1; estimated NVR=0.01



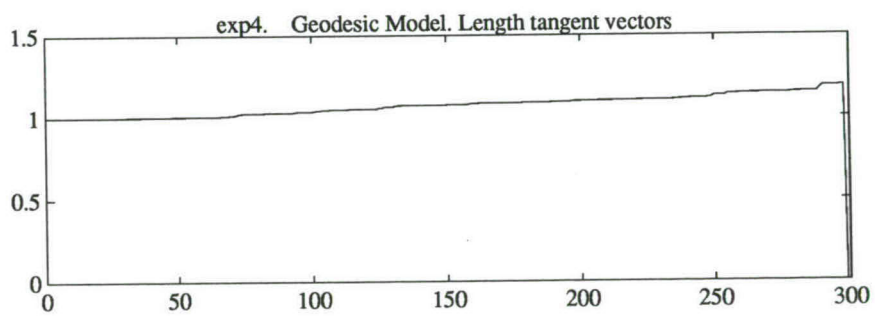
mean=-0.015; stand.dev.=0.294

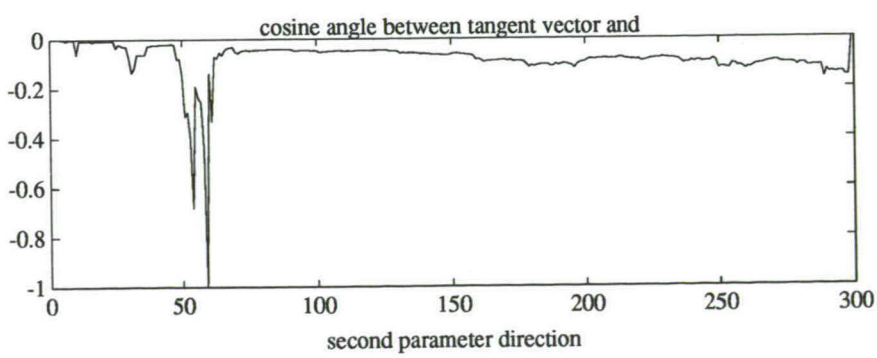


-2loglikelihood=116.3; estimated NVR=0.012

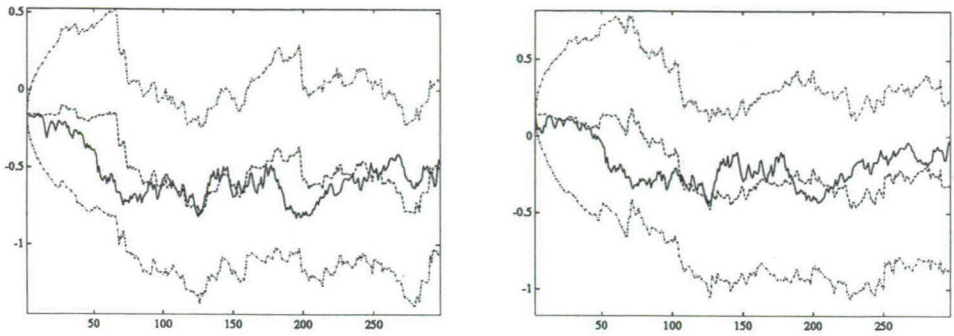


mean=-0.014; stand.dev.=0.284

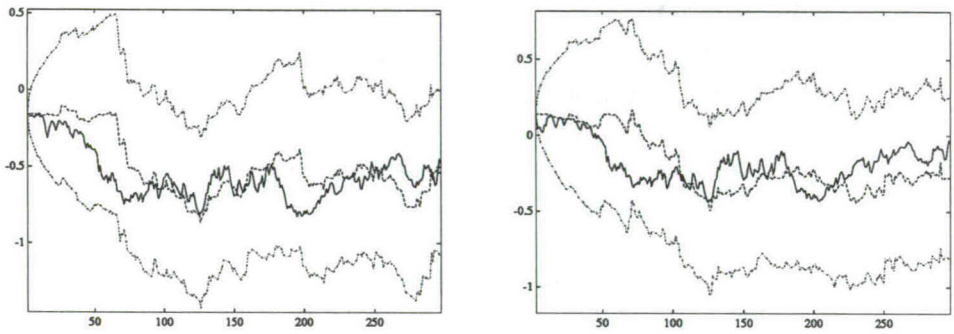




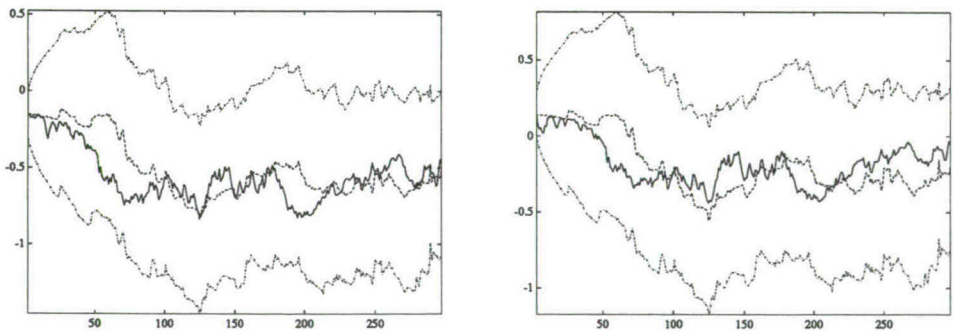
coefficient a_1 coefficient a_2



dashed: euclidean metric (Simple Model)



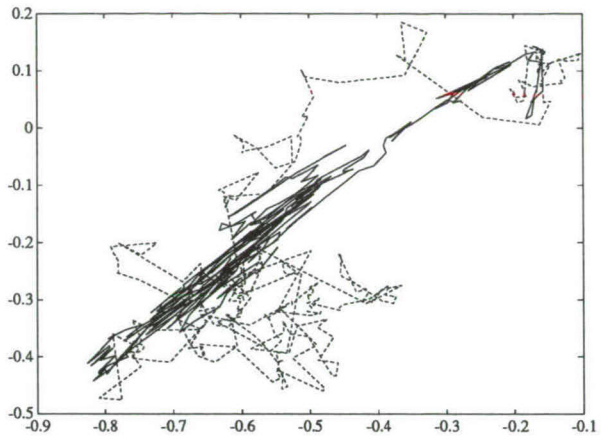
dashed: asymptotic Fisher metric (Simple Model)



dashed: asymptotic Fisher metric (Geodesic Model)

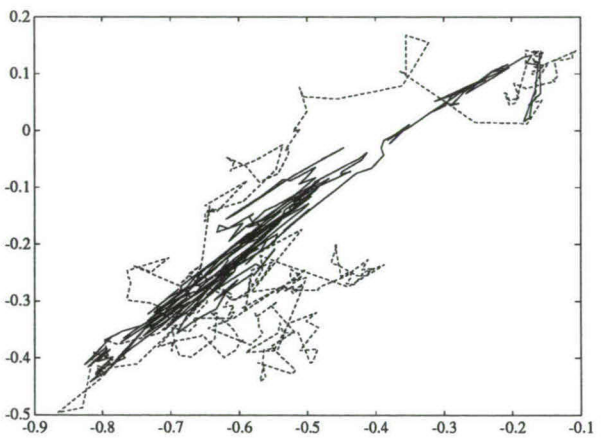
solid: generated coefficients dashdot: plus/min root estimated "error variance"

curve of coefficients a_1 and a_2

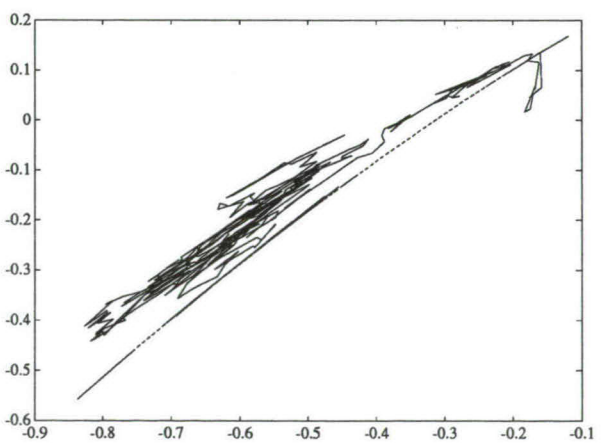


solid: curve of generated coefficients

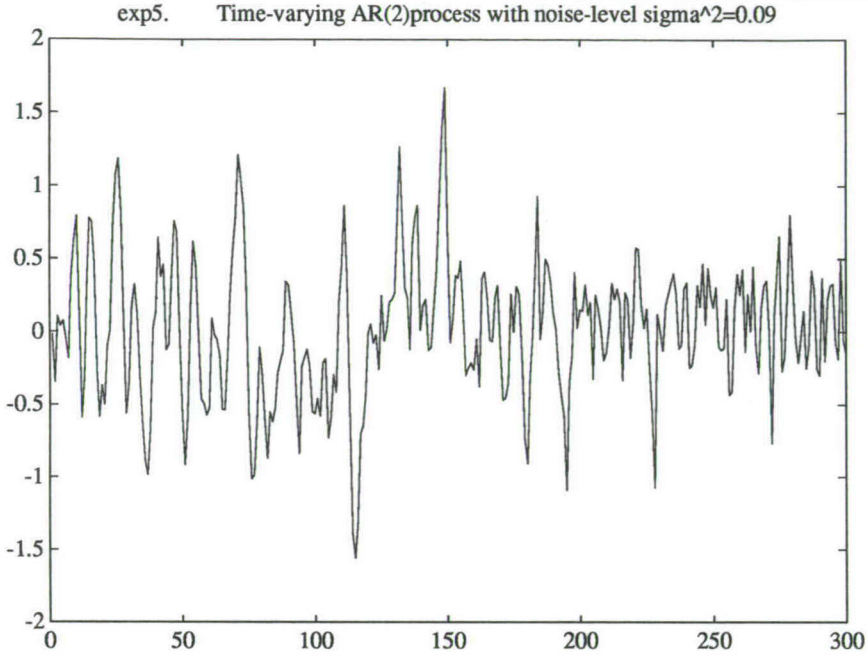
dashed: euclidean metric (Simple Model)



dashed: asymptotic Fisher metric (Simple Model)

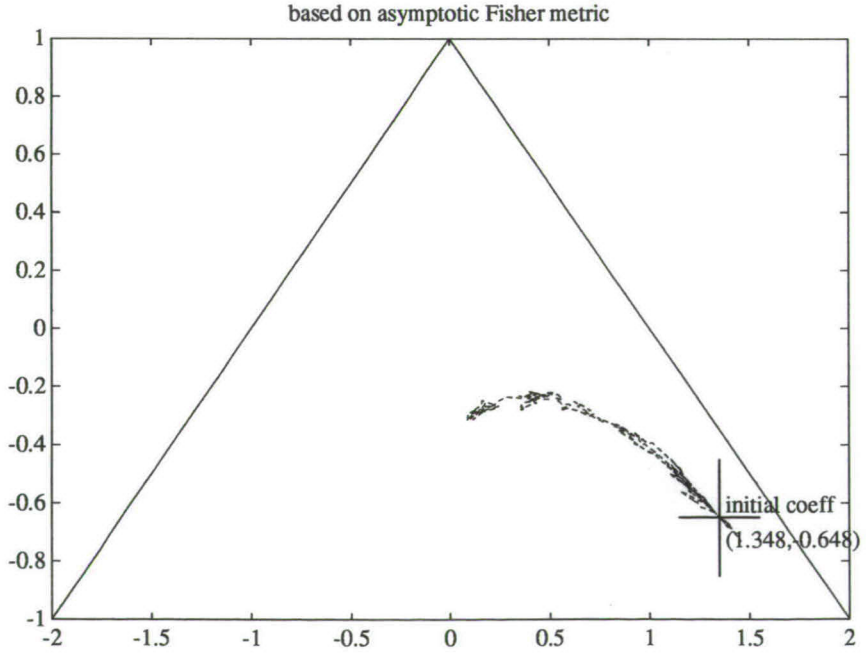


dashed: asymptotic Fisher metric (Geodesic Model)

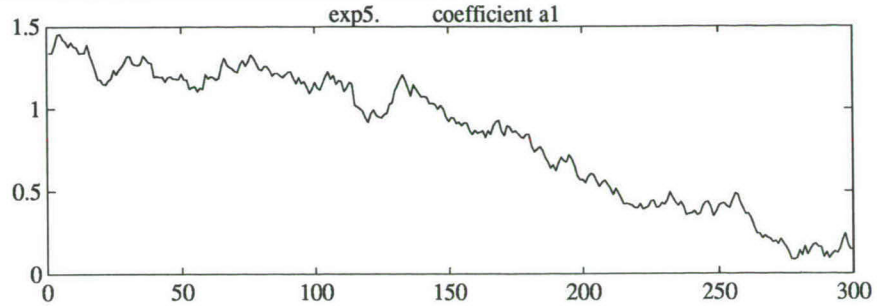


generated according to Special, Model of Complexity type 3

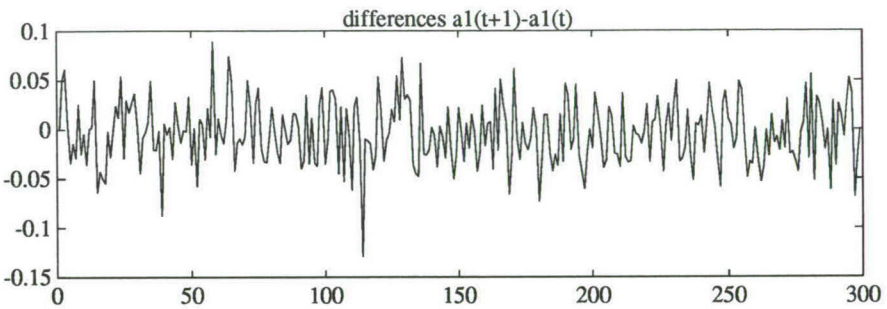
based on asymptotic Fisher metric



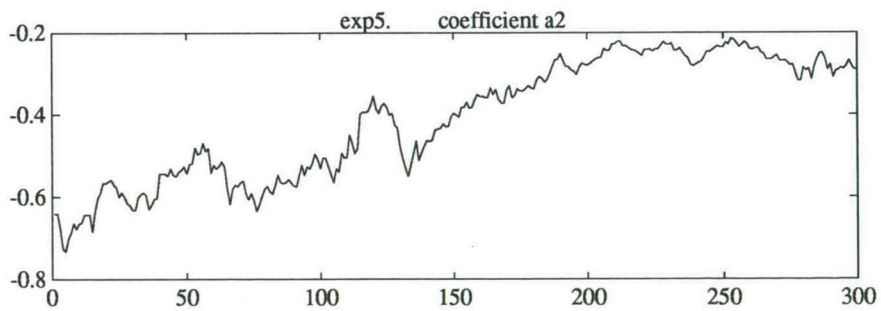
curve of coefficients



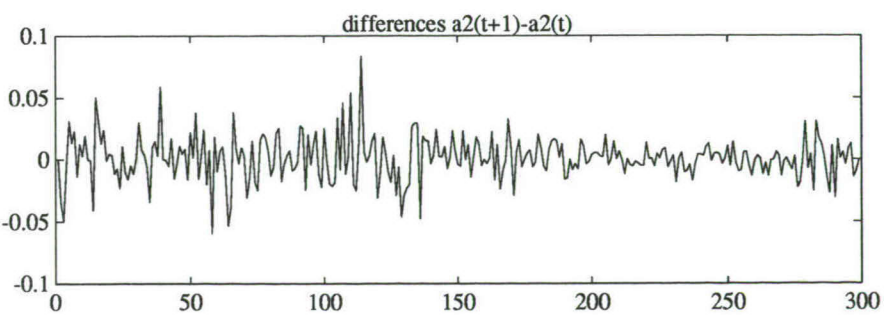
Special, Model of Complexity type 3 (asymptotic Fisher metric)



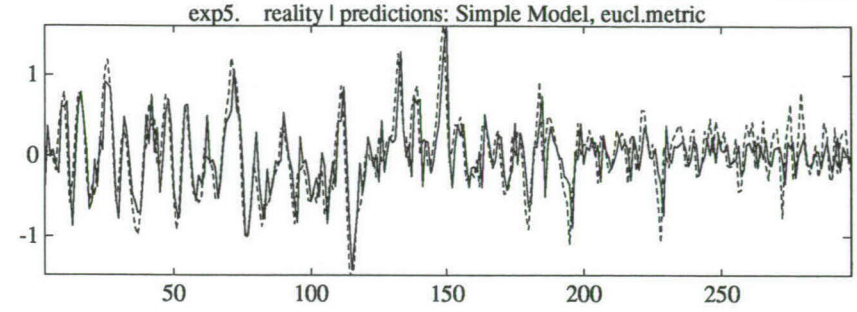
transition hyperparameters in jumps are tau1=0.008; tau2=0.04



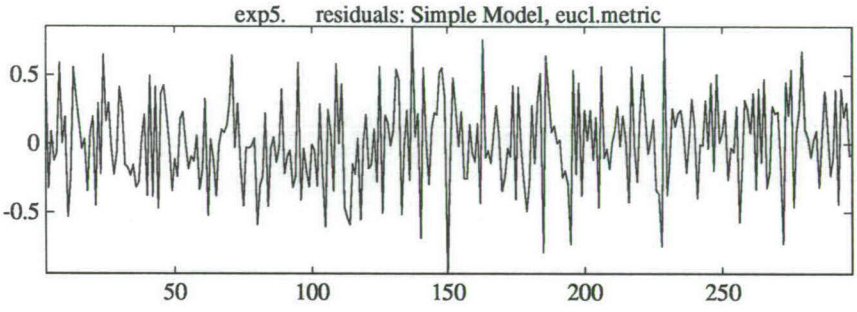
(300 jumps in coeff. on 300 observ.)



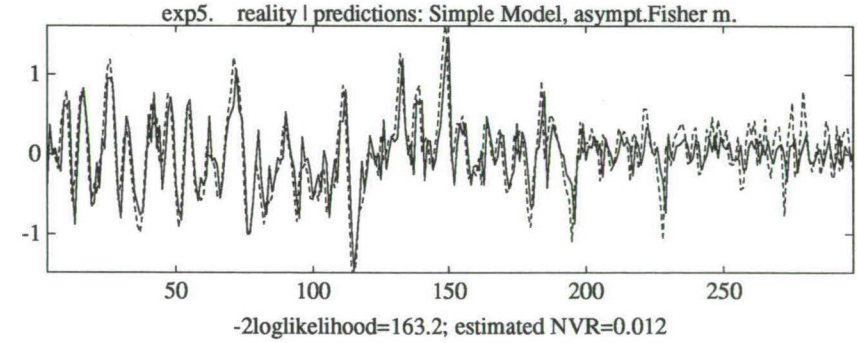
largest NVR=tau2^2/sigma^2=0.017 (Model of Complexity type 3!)



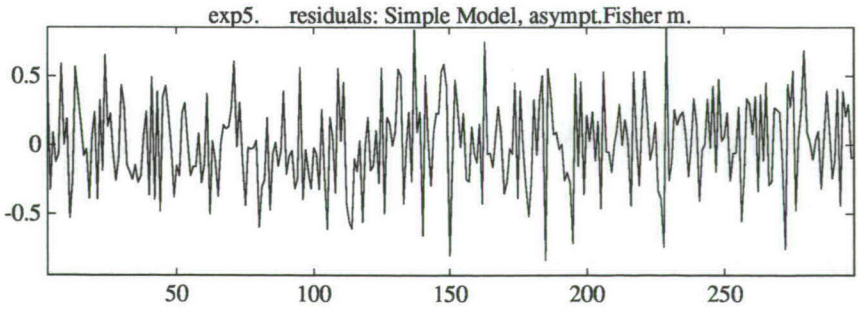
-2loglikelihood=166.4; estimated NVR=0.008



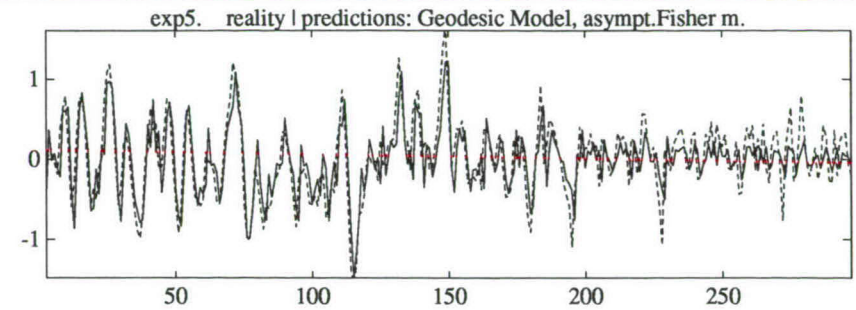
mean=0.013; stand.dev.=0.321



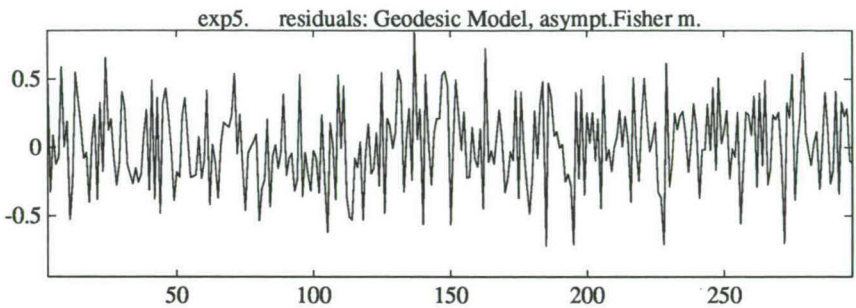
-2loglikelihood=163.2; estimated NVR=0.012



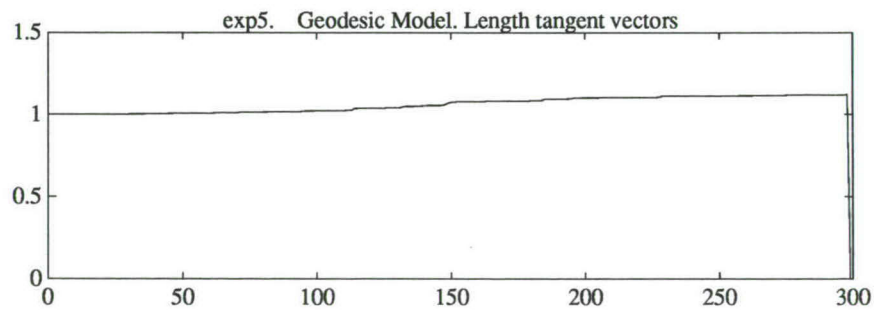
mean=0.013; stand.dev.=0.318

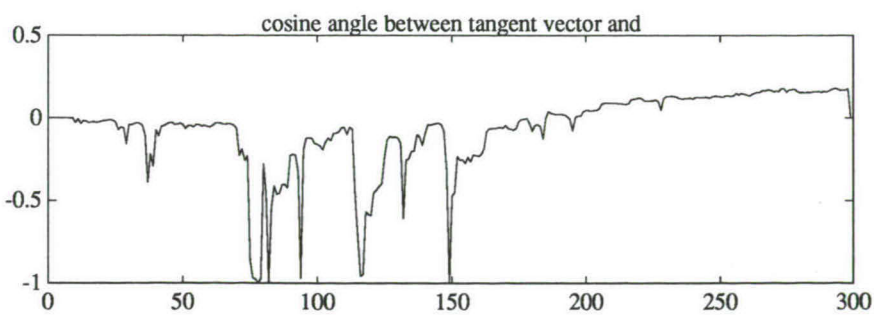


-2loglikelihood=159.6; estimated NVR=0.011



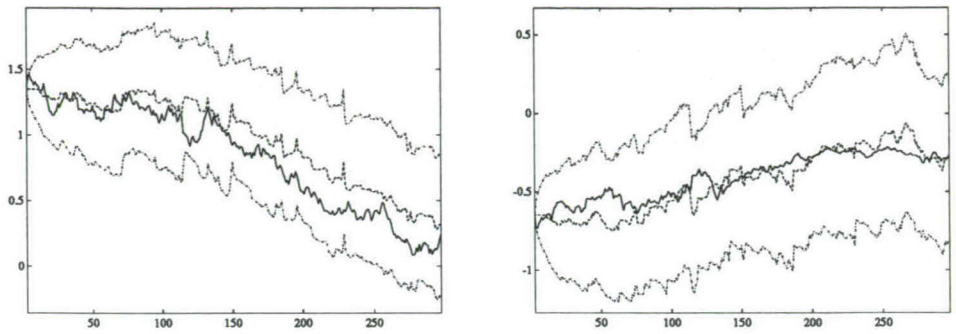
mean=0.019; stand.dev.=0.304



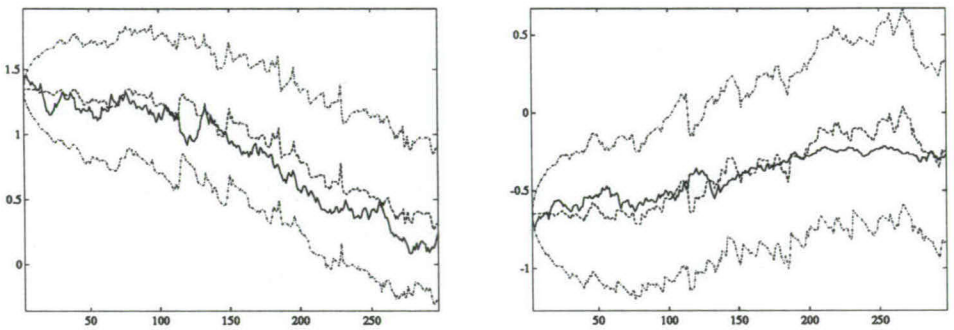


second parameter direction

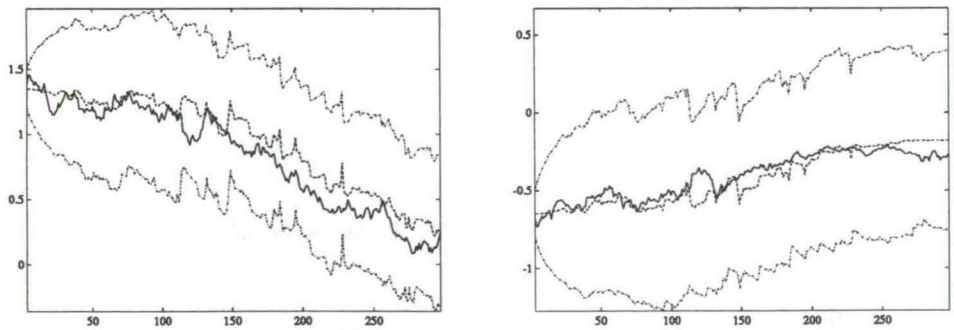
coefficient a_1 coefficient a_2



dashed: euclidean metric (Simple Model)



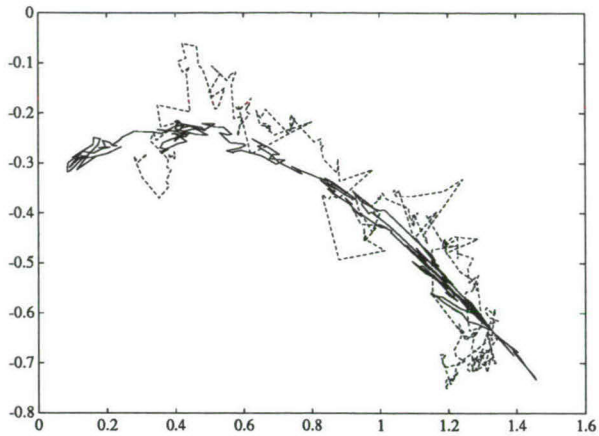
dashed: asymptotic Fisher metric (Simple Model)



dashed: asymptotic Fisher metric (Geodesic Model)

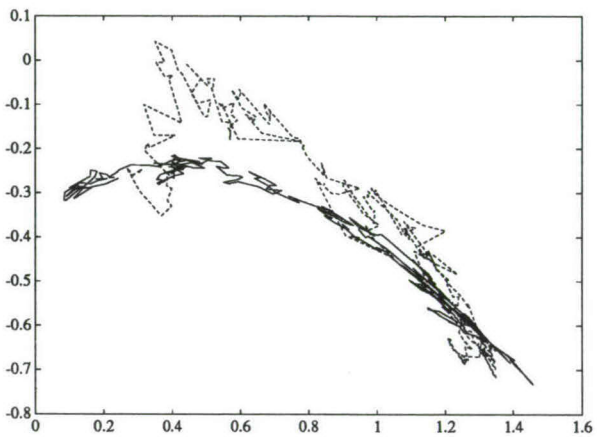
solid: generated coefficients dashdot: plus/min root estimated "error variance"

curve of coefficients a_1 and a_2

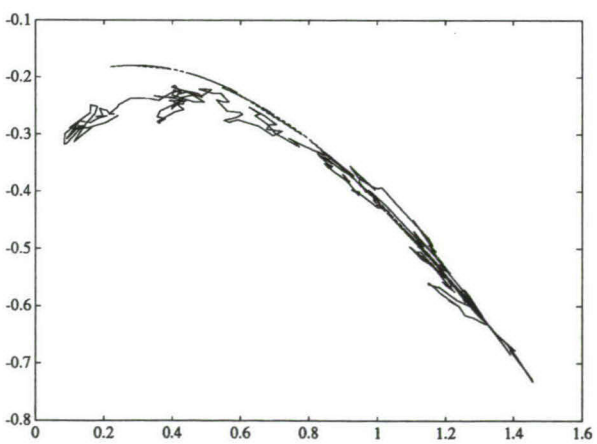


solid: curve of generated coefficients

dashed: euclidean metric (Simple Model)



dashed: asymptotic Fisher metric (Simple Model)



dashed: asymptotic Fisher metric (Geodesic Model)

Chapter 7

Conclusions

Models for time-varying coefficients of autoregressive processes can be compared with photographic material used in a camera. The model leads to a kind of photograph of the dynamics of the process which will enable us to predict the future of the dynamics and of the process itself. Just as one has films with different sensitivities (e.g. 21 DIN or 28 DIN) that have to be operated with different exposure times of the camera, we have models based on different metrics and with different types of direction spaces that have to be operated with different eigenvalues of the transition variance. The choice of metric, direction space etc. by the investigator can be compared with the choice of photographic material by a photographer. It depends on the situation (e.g. fast moving object in bright daylight or darkness), whether she opts for a film of 21 DIN or of 28 DIN. In order to make a conscious choice of a metric, the investigator has to know properties of the resulting models. In this thesis, mainly three different metrics and the corresponding models were studied: the Euclidean metric, complete metrics on U_n and the asymptotic Fisher metric on U_n. The Euclidean metric can be recommended, because of the computational ease when doing posterior mode estimation; an advantage of complete metrics on U_n is that resulting (sequences of) models have guaranteed stability properties, as elucidated in chapter four. Most interesting, but yet not completely investigated in all its consequences, is the stabilizing property of the asymptotic Fisher metric with respect to the observability gramian of the resulting model. We have encountered this metric in different situations. In the introduction (chapter one), we mentioned the case of a single jump in coefficients for an AR(1) process with constant noise level. If the asymptotic Fisher metric is used to model the jump, the transition variance eigenvalue can be interpreted as the expected Kullback-Leibler divergence between the time-invariant process before the jump and the time-invariant process after the jump. In chapter six, we mentioned the case of a sequence of geodesic models for autoregressive processes with fixed noise levels. If these geodesic models are based on the asymptotic Fisher metric, the expected contributions of all data to the information of the time-varying coordinates of the coefficients are equal asymptotically (in the Dahlhaus sense).
Both features (of the single-jump model and the geodesic models) are properties of zero-curvature models based on the asymptotic Fisher metric, and can be stated in the same way: the conditional expectation of the matrix of the observability gramian, given the coefficients, does not depend on these coefficients if a suitable parametrization of

the Model is chosen. It is with respect to this parametrization that also the matrix of the controllability gramian is constant on the coefficient space. For instance, reconsidering the example of the jumping AR(1) process in the introduction, we can use the parametrization $a = \sin(\kappa)$ $(a \in (-1,1), \kappa \in (-\frac{\pi}{2}, \frac{\pi}{2}))$. Then the model becomes:

$$\begin{aligned} \kappa_{t+1} &= \kappa_t + \tau_t \lambda_t \\ y_t &= \sin(\kappa_t) y_{t-1} + \epsilon_t \end{aligned} \tag{7.1}$$

where $\tau_t = 0$ except if $t = k$ (the jump-time). Linearizing the measurement equation in $\kappa_{t|t-1}$, we obtain:

$$\begin{aligned} \kappa_{t+1} &= \kappa_t + \tau_t \lambda_t \\ y_t - \sin(\kappa_{t|t-1}) y_{t-1} + \cos(\kappa_{t|t-1}) \kappa_{t|t-1} y_{t-1} &= \kappa_t \cos(\kappa_{t|t-1}) y_{t-1} + \epsilon_t. \end{aligned} \tag{7.2}$$

The controllability gramian for time $t = k$ is τ_k^2 and the expectation of the observability gramian (of the linearized system) for the same time instant is approximately one[1]; here, both the values τ_k^2 and 1 do not depend on the actual values of κ_k.

This property that the matrix of the controllability gramian *and* the expectation of the matrix of the observability gramian are constant on the coefficient space with respect to the same parametrization, enables us to derive easily uniform bounds for both matrices. For the geodesic model, we have seen that these bounds on their turn help in finding e.g. a positive lower bound for the gain matrix of Gauss-Newton iteration in posterior mode estimation of the coordinates of the coefficients. In all this, the parametrization of the zero-curvature model with respect to which the transition equation is linear, is important. For other models, we locally have parametrizations with respect to which the transition equation is almost linear in the sense of lemma 3.12. It is an interesting open question how far then properties hold, like constancy on (part of) the coefficient space of the matrix of both controllability gramian *and* expected observability gramian if the model is based on the asymptotic Fisher metric. The answer to this question can be helpful in explaining the difference in behaviour that we saw in the experiments of models based on Euclidean and asymptotic Fisher metric, especially if coefficients are close to the border of the stability region.

In all the identification methods presented in this thesis, we only used a single chart on the coefficient space. We already noted in the last section of chapter six that it is worth investigating whether for non-zero-curvature models, instead of a single chart, a *chain* of normal charts can be used with respect to which transition equations at different times are almost linear in the sense of lemma 3.12.

We have stressed that the choice of a metric implies an a priori view of the distances in the coefficient space; in other words, by choosing a metric we have given weights to all

[1]It is possible to give a coordinate free definition of the observability gramian and also of the controllability gramian for a General Model, for one and more time instants. Using such a coordinate free definition, the conditional expectation of the observability gramian for time $t = k$ given the coefficients, has a matrix with respect to the parametrization $a = \sin(\kappa)$, exactly equal to 1. A coordinate free definition for the observability gramian is possible, because the matrix of the observability gramian transforms in approximately the same way as the matrix H of the metric with coordinate transformations. One can conceive of the observability gramian as a kind of *posterior metric*.

possible coefficient changes. But in our models we have to make more a priori choices than just the selection of a metric. We also have to choose a direction space. We have elaborated simple and more complex models, where complexity depends on the dimension of the direction space.

A first stage in identifying the dynamics of an autoregressive process is the application of a simple model, i.e. a model with a direction space with a dimension equal to the dimension of the manifold (U_n or AR_n). A second, more refined stage is the application of a model, where the direction space has low dimension, e.g. dimension one. The lower the choice of dimension of the direction space, the more deterministic the coefficient coordinates become, and the more a mixture of Extended Kalman Filtering techniques and, e.g., Dahlhaus techniques becomes important. We have laid much emphasis on a typical example of such a model with low dimensional direction space; the geodesic model. In this model, the direction space $\mathcal{L}$ is one-dimensional and generates a totally geodesic distribution. There are more possibilities for models with a one-dimensional direction space $\mathcal{L} = \boldsymbol{span}\{X\}$. The spanning vector field X has constant length on the manifold by definition (see section 3.5 in chapter three). Let us say that this length is one. The integral curves of the vector field X are then parametrized, according to arc length. For the geodesic model the integral curves are geodesics.

In the thesis, we have mainly given one reason by which we more appreciate geodesic models than other models with a one-dimensional direction space. That reason is that the transition equation in the geodesic model is completely linear with respect to some chart, and hence posterior mode estimation of time-varying coefficient coordinates by Gauss-Newton procedures is possible. For other models with one-dimensional direction space, the linearization of the transition equation, as obtained in sections 3.6 and 3.7, can be completely linear, if a suitable chart is used, but this linearization is never completely equivalent to the true transition equation, according to proposition 3.9.

There is another, more qualitative property that make the geodesic model more interesting than other models with a one-dimensional direction space. It is a property of the "photographic material"-type. The property characterizes the distinguishability of coefficients with respect to each other. Once a distance measure on the coefficient space has a priori been chosen, a natural measure for "a priori expected distinguishability" of coefficient q_t from subsequent coefficients q_{t-1}, q_{t+1} is

$$\boldsymbol{E}(\boldsymbol{d}(q_{t-1}, q_t) + \boldsymbol{d}(q_t, q_{t+1})|q_{t-1}, q_{t+1}). \tag{7.3}$$

One can show that the geodesic model *maximizes* the a priori expected distinguishability of q_t if subsequent coefficients q_{t-1}, q_{t+1} are very close (if we restrict ourselves to a reasonable set of models with one-dimensional direction spaces).

The geodesic model becomes more flexible if the tangent field is varied with time, i.e. if the direction space is taken time-dependent: $\mathcal{L}_t = \boldsymbol{span}\{F_t\}$, F_t tangent fields of geodesics (in order to change the geodesic on which the coeficients are moving from time to time). We think that such time-varying geodesic models can e.g. be applied in electro-encephalogram-analysis, as a substitute for approaches in which simply chains of time-invariant autoregressive models are used. Geodesic models will smoothen these chains.

We presented a method to identify the Geodesic Model which is quite feasible, partic-

ularly by the fact that in this method, it is not necessary to know analytic formulae of the geodesics. Yet, we have to indicate some restrictions.
In the experiments and in the discussion of the hyperparameters, we have assumed that the direction space $\mathcal{L} = \boldsymbol{span}\{F\}$ is spanned by the tangent field of geodesics through a point q_{-1}, that was fixed, and known by the identifyer. The direction of the geodesic, i.e. the vector of $n-1$ hyperparameters ζ_0, the transition variance tensor eigenvalue τ^2 and the noise level σ^2 were left to be determined by maximum likelihood methods. Of course, the point q_{-1} gives n hyperparameters which can be determined by Gauss-Newton methods to maximize the likelihood in the same way as ζ_0, but the combined problem is rather complex, hence is only feasible for small n. It should be noticed that the approximation method for the parametrization, given in section 6.2, is also only feasible for small n due to the sharp increase in number of Christoffel- and curvature symbols that has to be computed for larger n.
We note that the presented method has been implemented and tested out experimentally if measurement noise levels are fixed or non-stochastic. In order to include stochastically time-varying noise levels, we introduced the Geodesic plus Noise model with two-dimensional direction space. We have indicated how this model can be identified. These identification methods should be tested out yet.
We think that the Geodesic Model is attractive as a refinement of a low-order Simple Model, as an alternative to increasing the order of the Simple Model. It is an alternative in which the model is made more deterministic in some sense. This is clear if one looks at the proportion between the number of hyperparameters involved, and the coordinates that are still to be estimated by Kalman methods. The geodesic model (certainly the special geodesic model) has a scalar transition equation, i.e. on some charts there is only one coordinate changing with time; the others are fixed and can be regarded as hyperparameters. Hence, in formulating the geodesic model, we are already shifting emphasis considerably from stochastic modelling towards deterministic modelling. This comes out strongly in the adoption of the asymptotics of Dahlhaus. Penalties in the log-joint-density-expression in this asymptotics go to infinity; only coefficients coordinates with intervals proportional to $T^{\frac{3}{4}}$ can asymptotically be determined by a Kalman Filter (if the asymptotic Fisher metric is used, even by a time-invariant filter). Hence, asymptotically, the main job is to determine the hyperparameters.

Our interest in some kind of asymptotics is related to the fact that, by the introduction of low-dimensional direction spaces, we had to face the problem of estimating deterministic coordinates, like the direction of the geodesic through q_{-1}, ζ_0, in the geodesic model. We do not only have to estimate ζ_0, but also the error variance of the estimator of ζ_0. This problem has not been completely solved in this thesis. Yet, the way to its solution is completely clear.
We have given a reasonable approximation method for the *error variance* for the estimates of the stochastic moving coordinates: just the diagonal elements of the inverse of the gain matrix U in Fahrmeir/Kaufmann's method. The error variance of estimates of the geodesic direction ζ_0 can be given in two ways:

- if ζ_0 is regarded as an ordinary parameter, the inverse of the gain matrix U has a diagonal element that can be seen as an indication for the error variance of ζ_0.

We have seen that it is difficult to work with this matrix U (and to compute its inverse, for instance) because the last column and last row are non-zero everywhere (this column and row contain the entries corresponding to the double derivatives $\frac{\partial^2}{\partial\zeta\partial\kappa_t}$). However, we conjecture that asymptotically (coefficients being estimated with intervals proportional to $T^{\frac{3}{4}}$) the gain matrix U becomes pentadiagonal also in this case (i.e. the last column and last row become zero too) if the asymptotic Fisher metric is used.

- if ζ_0 is regarded as hyperparameter, the error variance can be approximated with the asymptotic distribution of its MLE-estimator. It can be investigated how far the results of Dahlhaus can be used for this.

It is interesting to investigate whether, asymptotically, there is any difference between the two estimates: the MAP estimate of ζ_0 together with $\kappa_{0,T}, \kappa_{sT^{\frac{3}{4}},T}, \kappa_{2sT^{\frac{3}{4}},T}, \cdots$ and the ML estimate of ζ_0, in case the model is based on the asymptotic Fisher metric. We conjecture that there is none.

To end this chapter, we state the main results obtained in the different chapters.

Chapter three We have generalized the random walk /smoothly integrated random walk model for a coefficient process to the Special /General Model of a polygonal process of independent directional variation on an arbitrary Riemannian manifold. These generalizations take account of specific geometric structures on the coefficient space. It implies an a priori weighting of coefficient changes in an appropriate, sensible way. We have given necessary and sufficient conditions such that there exist coordinates on the manifold and the polygonal coefficient process, expressed in these coordinates, is just a random walk/ smoothly integrated random walk.
The developed theory is not only important for time-varying autoregressive processes, but for all kinds of time-varying processes, certainly all the processes which Fahrmeir and Kaufmann want to model by Dynamic Exponential Family Regression. We think that properties of the Fisher metric in these other applications will be of the same kind as we found for the asymptotic Fisher metric for autoregressive processes.

Chapter four We have given arguments for our adoption of the Dahlhaus way of performing asymptotics. We have linked the existence of a time-varying spectral density for a sequence of models to the question of stability of this sequence. We have shown that the time-varying spectral density can be introduced for sequences of models of autoregressive processes, where coefficients are not deterministic, but follow smoothly integrated random walks.
For further research, we note that this time-varying spectral density could be the basis of an estimation method for the hyperparameters of the model, in line with the theory designed by Dahlhaus. Besides the estimate of the hyperparameter itself, this method could asymptotically provide an error variance.

Chapter five We have studied the Riemannian geometry on the stability region U_n for time-invariant stationary AR(n) processes given by the asymptotic Fisher metric for such processes. In particular, we studied covariance stabilizing parametrizations. We have shown that such a covariance stabilizing parametrization does not exist for the whole U_n if $n > 1$.

Chapter six We have shown the advantage that models with linear transition equation have above other models: the possibility to do posterior mode estimation by Gauss-Newton iterations, using the Fahrmeir/Kaufmann's method. We have adapted the Fahrmeir/Kaufmann-method to be applicable also if the transition equation has the form

$$\begin{aligned} \tilde{\kappa}_{t+1} &= \tilde{\kappa}_t + \tilde{C}\tilde{\theta}_t \\ \tilde{\theta}_{t+1} &= \tilde{A}\tilde{\theta}_t + \tilde{\lambda}_t, \end{aligned} \tag{7.4}$$

where $(\tilde{\lambda}_t)_{t\in\mathbb{Z}}$ is Gaussian white noise.
We have shown that for the identification of geodesic models it is not necessary to have analytic formulae for the geodesics.
We have developed a quite feasible method of identifying geodesic models with fixed, non-stochastic noise-levels.
We have experimentally shown that this method works well and gives interesting results, when compared with the results of more usual models.
We have indicated how to identify models, where also the noise-levels are stochastically modelled.
For further research, we note that for non-zero-curvature models, one could study the use of chains of normal charts instead of using one single chart for the identification.
As an alternative to ordinary maximum likelihood estimation, we have already recommended to investigate the use of the time-varying spectral density in order to estimate hyperparameters, and to compute the error variance of such an estimate asymptotically. Not all hyperparameters can be estimated using the time-varying spectral density, because this density is conditional on an Ornstein-Uhlenbeck Velocity process, hence one should determine the hyperparameters of this OUV process in another way. Also if we are not able to do asymptotics (like in a Special Model), we cannot rely on this method in order to estimate the hyperparameters. It is interesting to investigate, whether in this situation also the idea can be followed of Young, Ng and Armitage (1989), and another type of spectral techniques (based on Kalman and Wiener/Kolmogorov filters) can be used to get a first estimate of a hyperparameter after which more usual maximum likelihood methods are applied.

Appendix 1 We show that it is possible to give coordinate free definitions of the controllability gramian and observability gramian for a Special and General Model. We used this in the conclusions to formulate properties of zero-curvature models based on the asymptotic Fisher metric in a more unified way. We restrict ourselves to Special Models of autoregressive processes with fixed noise levels.
For a Special Model, the controllability gramian $\mathcal{K}(t,t)$ for one time instant t is just the transition variance, and the transition variance was already defined in a coordinate free way in chapter two. For a Special Model with direction space $\mathcal{L}$, the definition becomes

$$\mathcal{K}(t,t)(X,Y) = \langle V_t(X(q_t)), Y(q_t)\rangle_{q_t} \quad (X,Y \in \mathcal{L}). \tag{.5}$$

By proposition 3.6 we know that the expression on the right hand side does not depend on the value of q_t, i.e.

$$\langle V_t(X(q_t)), Y(q_t)\rangle_{q_t} = \langle V_t(X(p)), Y(p)\rangle_p \text{ for all } t \in \mathbf{Z},\ p \in M, X, Y \in \mathcal{L}, \tag{.6}$$

hence it is not necessary to indicate the argument q_t (we just use the inner product of the direction space itself, see proposition 3.6). The controllability gramian for two subsequent time instants $t-1, t$ can be defined in a natural way by

$$\mathcal{K}(t,t-1)(X,Y) = \mathcal{K}(t,t)(X,Y) + \langle V_{t-1}(par(X(q_t))), par(Y(q_t))\rangle_{q_{t-1}} \quad (X,Y \in \mathcal{L}). \tag{.7}$$

Here, $par(v)$ is the parallel shift of the tangent vector v in the tangent space M_{q_t} to the tangent space $M_{q_{t-1}}$ along the geodesic connecting q_t and q_{t-1}. If $\mathcal{L}$ satisfies the property which guarantees the existence of a parametrization with respect to which there is a linear transition equation, i.e. $\nabla_X Y = 0$ for all X, $Y \in \mathcal{L}$, then the parallel shift is just $par(X(q_t)) = X(q_{t-1})$ for all $X \in \mathcal{L}$. In this case, the controllability gramian for N subsequent time instants $t+1, \cdots, t+N$ for a Special Model with one direction space $\mathcal{L}$ will be

$$\mathcal{K}(t+N,t+1)(X,Y) = \langle \sum_{j=1}^{N} V_{t+j}(X), Y\rangle = \sum_{j=1}^{N} \mathcal{K}(t+j,t+j)(X,Y) \quad (X,Y \in \mathcal{L}). \tag{.8}$$

Also the observability gramian depends on the parametrization or chart used for the coefficient space. However, the observability gramian $\mathcal{J}(t,t)$ at one time instant t transforms in approximately the same way as the matrix H of the metric with coordinate transformations. This can be shown clearly for a Special Model on the manifold $\frac{1}{\sigma}AR_n$. In a Special Model, a (linearized) transition equation has the form:

$$a_{t+1} = a_t + g(a_t)C_t\lambda_t \tag{.9}$$

and the measurement equation is:

$$y_t = \phi_t' a_t + \sigma\epsilon_t. \tag{.10}$$

The observability gramian for one time instant t is $\frac{\phi_t\phi_t'}{\sigma^2}$. If we use the parametrization $\psi(\mu) = (a_1, \cdots, a_n)$ and linearize the measurement equation in the point $(\mu_{t|t-1})_{t\in\mathbf{N}}$, this measurement equation becomes:

$$y_t - \phi_t'\psi(\mu_{t|t-1}) + \phi_t' D\psi(\mu_{t|t-1})\mu_{|t-1} = \phi_t' D\psi(\mu_{t|t-1})\mu_t + \sigma\epsilon_t, \tag{.11}$$

and accordingly the observability gramian for one time instant t now is $D\psi'(\mu_{t|t-1})\frac{\phi_t\phi_t'}{\sigma^2}D\psi(\mu_{t|t-1})$. One can compare this transformation with the transformation of the matrix of the metric H_{id} into H_ψ, see chapter two, equation (2.4). This fact suggests an attempt to define also the observability gramian in a coordinate free way. One could think of the observability gramian for a Special Model as kind of a *posterior metric.*

The observability gramian for one time instant t, $\mathcal{J}(t,t)$, for a Special Model with direction space $\mathcal{L}$ and constant measurement noise level σ^2 will be

$$\mathcal{J}(t,t)(X,Y) = \sum_{i,j=1}^{n} X(q_t)_i \frac{y_{t-i}y_{t-j}}{\sigma^2} Y(q_t)_j. \quad (X,Y \in \mathcal{L}\,) \tag{.12}$$

if $X(q_t) = \frac{1}{\sigma}(X(q_t)_1 x^{n-1} + \cdots + X(q_t)_n) \in (\frac{1}{\sigma}AR_n)_{q_t}$. In the same way as above, one can use parallel shifts in order to define the observability gramian for time instants $t+1, \cdots, t+N$.

It is easy to verify that the conditional expectation of the observability gramian for one time instant t given the coefficients *in case of a single jump at time* k has the following form:

$$\boldsymbol{E}^{\mathcal{B}_T}\mathcal{J}(t,t)(X,Y) = \langle X(q_t), Y(q_t)\rangle_{q_t,\text{asympt.Fisher metric}} \text{ for all } t \leq k, X, Y \in \mathcal{L}. \tag{.13}$$

If the single jump model is also based on this asymptotic Fisher metric, then the expectation of the observability gramian at one time instant is just equal to this a priori metric, until the jump, and does not depend on the actual value of q_t, just as the controllability gramian. If we have a parametrization for the model with respect to which the transition equation is linear, then the controllability gramian for one time instant has a matrix constant on the coefficient space and then also the expectation of the observability gramian is just the identity matrix. For instance, in model (7.1), the controllability gramian for time $t = k$ is τ_k^2 and the expected observability gramian is 1 with respect to the parametrization $a = \sin(\kappa)$, and does not depend on the actual value of κ_k. Hence, for one parametrization we then have the matrix of the controllability gramian constant on the coefficient space *and* the matrix of the conditional expectation of the observability gramian constant on the coefficient space. Asymptotically, we have seen the same feature for geodesic models based on the asymptotic Fisher metric in chapter six.

Appendix 2 We show a claim about a property of the geodesic model we mentioned in the conclusions but that was not integrated in the other chapters because it did not fit there. It is the property of maximal a priori expected distinguishability of a coefficient between two others if these others are very close.

We are going to compare Special Models on the same manifold M with the same metric, with a one-dimensional direction space $\mathcal{L} = \boldsymbol{span}\{X\}$ and with the same transition variance tensor eigenvalue τ^2. As we have seen in section 3.5 of chapter three, the vector field X has constant length on the manifold by definition. We may assume that this length $||X||$ is equal to one. Consequently, any integral curve $\kappa \mapsto f_X(\kappa)$ of this vector field is parametrized according to arc length. For a model with direction space

$\mathcal{L} = \boldsymbol{span}\{X\}$ one can cover the manifold M with charts $(\tilde{U}_\alpha, x_\alpha)$ such that $\frac{\partial}{\partial x_{\alpha\ n}} = X$. (Flow-box theorem). Linearizing the transition equation of the special model on such a chart yields

$$\begin{aligned} \kappa_{t+1} &= \kappa_t + \tau\lambda_t \text{ where} \\ \lambda_t &\overset{i.i.d}{=} \mathcal{N}(0,1) \text{ for all } t \text{ and} \\ q_t &= f_X(\kappa_t). \end{aligned} \tag{.14}$$

We compare these models with respect to the expected distinguishability of the generated coefficients. Once we have a priori chosen for a distance measure on the coefficient space, a natural measure for distinguishability between three subsequent coefficients q_0, q_1, q_2 is

$$\boldsymbol{d}(q_0, q_1) + \boldsymbol{d}(q_1, q_2). \tag{.15}$$

The models generate the coefficients stochastically, hence expression (.15)is a $\mathcal{A}-$ or even a $\mathcal{B}_2$-measurable-function. A criterion of optimality for distinguishability is maximization of the conditional expectation of expression (.15), given q_0 and q_2, if these coefficients are closer than a certain number R. The maximization should be over a model family such that

- the family only contains models for which the vector field X spanning the direction space, has integral curve f_X through the points q_0 and q_2;

- the family also contains the geodesic model through q_0 and q_2, i.e. the model such that its direction space has an integral curve f_X through the points q_0 and q_2 and such that this integral curve is a geodesic;

- between these points q_0 and q_2, the family of models is "not too far" from the geodesic model through q_0 and q_2
  (the notion "not too far" will be quantitively interpreted below);

- far from these points q_0 and q_2, the family of models is "reasonably far" from the geodesic model through q_0 and q_2
  (The notion "reasonably far" will also be quantificated below).

We can show that for small R, the geodesic model through q_0 and q_2 constitutes a (local) maximum of the conditional expectation of expression (.15) over this family of models. We consider subfamilies of this family of models which can be described (or parametrized) by giving the corresponding integral curves f_X through q_0 and q_2 as follows:

$$f_X(\kappa) = f(z, \kappa) \quad (z \in W \subset \mathbb{R},\ \kappa \in I \subset \mathbb{R}) \tag{.16}$$

where

- the function $f : W \times I \to M$ is C^∞;

- $f(z, -\frac{1}{2}L_z) = q_0$; $f(z, \frac{1}{2}L_z) = q_2$ for some positive number L_z for all $z \in W$.;
- $f(z,\kappa) \neq q_0, \neq q_2$ for all κ, $\kappa \neq -\frac{1}{2}L_z, \frac{1}{2}L_z$;
- $||\frac{\partial f}{\partial \kappa}(z,\kappa)||_{f(z,\kappa)} = 1$ (i.e. the curves are parametrized according to arc length)
- and the curve $\kappa \mapsto f(z,\kappa)$ is the geodesic through q_0 and q_2 if and only if $z = 0$.

In this setup, the length of the curve $\kappa \mapsto f(z,\kappa)$ between q_0 and q_2 is L_z. If R is small enough, the distance between q_0 and q_2 is equal to the length of the geodesic, hence

$$d(q_0, q_2) = L_0 \text{ and } L_z > L_0 \text{ for all } z \neq 0. \tag{.17}$$

Consequently, we have

$$\frac{\partial}{\partial z} L_{z|z=0} = 0 \text{ and } \frac{\partial^2}{\partial z^2} L_{z|z=0} \geq 0. \tag{.18}$$

We shall say that between the points q_0 and q_2, the family of models is *not more than η-far* from the geodesic model through these points, if

$$\frac{\partial^2}{\partial z^2} L_{z|z=0} \leq \eta. \tag{.19}$$

The distance of a point $f(z,\rho)$ to the point q_2 is shorter than the length of the curve $\kappa \mapsto f(z,\kappa)$ between these points, hence shorter than $|\frac{1}{2}L_z - \rho|$. It is equal to this length if and only if the curve is a geodesic, hence if and only if $z = 0$. We introduce the function

$$v_2(z,\rho) = |\frac{1}{2}L_z - \rho| - d(f(z,\rho), q_2). \tag{.20}$$

Then

$$v_2(z,\rho) \geq 0 \text{ for all } z, \rho \text{ and } v_2(0,\rho) = 0 \text{ for all } \rho \in I. \tag{.21}$$

Consequently,

$$\frac{\partial}{\partial z} v_2(z,\rho)_{|z=0} = 0 \text{ and } \frac{\partial^2}{\partial z^2} v_2(z,\rho)_{|z=0} \geq 0. \tag{.22}$$

Following the same idea, we also introduce the function

$$v_0(z,\rho) = |\frac{1}{2}L_z + \rho| - d(f(z,\rho), q_0). \tag{.23}$$

Then

$$v_0(z,\rho) \geq 0 \text{ for all } z \in W, \rho \in I \text{ and } v_0(0,\rho) = 0 \text{ for all } \rho \in I. \tag{.24}$$

Also here, we conclude

$$\frac{\partial}{\partial z} v_0(z,\rho)_{|z=0} = 0 \text{ and } \frac{\partial^2}{\partial z^2} v_0(z,\rho)_{|z=0} \geq 0. \tag{.25}$$

We shall say that far from the points q_0 and q_2, the family of models is *more than ϵ-far* from the geodesic model through these points, if

$$\begin{aligned} \frac{\partial^2}{\partial z^2} v_2(z,\rho)_{|z=0} &\geq \epsilon \text{ if } \rho > L_0 \text{ and} \\ \frac{\partial^2}{\partial z^2} v_0(z,\rho)_{|z=0} &\geq \epsilon \text{ if } \rho < -L_0. \end{aligned} \tag{.26}$$

For a family that, between the points q_0 and q_2, is not more than η-far from the geodesic model through these points, and, far from these points, is more than ϵ-far, we shall prove that

$$\boldsymbol{E}(\boldsymbol{d}(q_0,q_1)+\boldsymbol{d}(q_1,q_2)|q_0,q_2) \tag{.27}$$

attains its maximum in the geodesic model through q_0 and q_2 if the distance of these points is small enough (i.e. if R is small enough). We define the function $g: W \times I \to \mathbb{R}^+$ by

$$g(z,\kappa) = \boldsymbol{d}(f(z,\kappa),q_0) + \boldsymbol{d}(f(z,\kappa),q_2). \tag{.28}$$

We want to maximize the quantity

$$K(z) := \boldsymbol{E}(g(z,\kappa_1)|\kappa_0 = -\frac{1}{2}L_z; \kappa_2 = \frac{1}{2}L_z; \kappa_1 \in I) \tag{.29}$$

as a function of z.

Statement. $K(z)$ has a stationary point in $z=0$. If q_0 and q_2 are sufficiently close (i.e. if R is small enough), then $K(z)$ has a local maximum in $z=0$.

Proof: From equations (.14) one concludes that the distribution of κ_2 given $\kappa_0 = -\frac{1}{2}L_z$ is normal with mean $-\frac{1}{2}L_z$ and variance equal to $2\tau^2$. Now, it is easily derived (using Bayes' rule) that the conditional distribution of κ_1 given $\kappa_0 = -\frac{1}{2}L_z$ and $\kappa_2 = \frac{1}{2}L_z$ is normal with zero mean and variance $\frac{1}{2}\tau^2$. Furthermore, we take the distribution of κ_1 conditional on the event $\kappa_1 \in I$. Then the conditional probability density becomes:

$$p(x) := \frac{e^{-\frac{x^2}{\tau^2}}}{\int_I e^{-\frac{x^2}{\tau^2}} dx} 1_I(x). \tag{.30}$$

Then we can write:

$$K(z) = \int_I g(z,\kappa)p(\kappa)d\kappa. \tag{.31}$$

We study the derivatives of g with respect to z. Note that

$$L_0 \leq g(z,\kappa) \leq |\frac{1}{2}L_z - \kappa| + |\frac{1}{2}L_z + \kappa| \text{ for all } z \in W, \kappa \in I. \tag{.32}$$

For κ satisfying $|\kappa| < \frac{1}{2}L_0$ we have:

$$L_0 \leq g(z,\kappa) \leq L_z, \text{ equality signs if } z = 0. \tag{.33}$$

Consequently, with (.18) and (.19) we have

$$\frac{\partial}{\partial z} g(z,\kappa)_{|z=0} = 0 \text{ and } 0 \leq \frac{\partial^2}{\partial z^2} g(z,\kappa)_{|z=0} \leq \frac{\partial^2}{\partial z^2} L_{z|z=0} \leq \eta \text{ for all } \kappa,\ |\kappa| < \frac{1}{2}L_0. \tag{.34}$$

For κ satisfying $|\kappa| > \frac{1}{2}L_0$ and z sufficiently close to 0 we have:

$$|\kappa| > \frac{1}{2}L_z \text{ hence } g(z,\kappa) \leq 2|\kappa|, \text{ and } g(0,\kappa) = 2|\kappa| \tag{.35}$$

accordingly,

$$\frac{\partial}{\partial z}g(z,\kappa)_{|z=0} = 0 \text{ and } \frac{\partial^2}{\partial z^2}g(z,\kappa)_{|z=0} \leq 0 \text{ for all } \kappa,\ |\kappa| > \frac{1}{2}L_0. \tag{.36}$$

We can sharpen the inequality (.36) with (.26). For κ satisfying $|\kappa| > \frac{1}{2}L_0$ (and z sufficiently close to 0) we have

$$g(z,\kappa) - 2|\kappa| = -v_0(z,\kappa) - v_2(z,\kappa). \tag{.37}$$

Consequently,

$$\frac{\partial^2}{\partial z^2}g(z,\kappa)_{|z=0} = -\frac{\partial^2}{\partial z^2}v_0(z,\kappa)_{|z=0} - \frac{\partial^2}{\partial z^2}v_2(z,\kappa)_{|z=0} \leq -\epsilon \text{ if } |\kappa| > L_0. \tag{.38}$$

With these (in)equalities it is easy to compute the derivative of $z \mapsto K(z)$ in $z = 0$:

$$K'(0) = \int_I \frac{\partial}{\partial z}g(z,\kappa)p(\kappa)d\kappa = 0. \tag{.39}$$

We can also derive an inequality for the second derivative:

$$K''(0) = \int_I \frac{\partial^2}{\partial z^2}g(z,\kappa)p(\kappa)d\kappa \leq \eta\int_{-\frac{1}{2}L_0}^{\frac{1}{2}L_0} p(\kappa)d\kappa - \epsilon\int_{L_0}^{\infty} p(\kappa)d\kappa - \epsilon\int_{-\infty}^{-L_0} p(\kappa)d\kappa. \tag{.40}$$

It is clear that if L_0 is small enough, $K''(0)$ will be smaller than zero and $K(z)$ will have a local maximum in zero. □

Although the result is rather limited, it shows that if we have chosen a metric for its good observability properties, and we have to decide on a model with one-dimensional direction space, a natural choice will be to take models with direction space $\mathcal{L} = \boldsymbol{span}\{F\}$, F a tangent field of geodesics.

Bibliography

[1] H.Akaike, *A new look at statistical model identification,* In: IEEE Trans. in Autom. Control, vol AC-19, no 6, 1974, pp. 716-722.

[2] S.-I.Amari, *Differential Geometry of Curved Exponential Families-Curvatures and Information Loss,* In: The Annals of Statistics, vol. 10, no. 2, 1982, pp. 357-385.

[3] S.-I.Amari, *Differential Geometry of a Parametric Family of Invertible Linear Systems-Riemannion Metric, Dual Affine Connections, and Divergence,* In: Math. Systems Theory, vol 20, 1987, pp. 53-82.

[4] S.-I.Amari, *Differential-Geometrical Methods in Statistics,* 2nd edition 1990, Springer-Verlag.

[5] B.D.O. Anderson and J.B. Moore, *Optimal Filtering,* 1979, Prentice-Hall.

[6] M.Arato, *On the sufficient statistics for stationary gaussian random processes,* In:Theory of Probability and its Applications, Vol. VI, No. 2, 1961, pp. 199-201.

[7] O.Barndorff-Nielsen and G.Schou, *On the Parametrization of Autoregressive Models by Partial Autocorrelations,* In: Journal of Multivariate Analysis, Vol. 3, 1973, pp. 408-419.

[8] K.N.Berk, *Consistent autoregressive spectral estimates,* In: Annals of Statistics, Vol. 2, no. 3, 1974, pp.489-502.

[9] G.E.P.Box and G.M.Jenkins, *Time Series Analysis-forecasting and control,* Revised edition 1976, Holden-Day, San Francisco.

[10] R.A.Boyles, *On the Convergence of the EM Algorithm,* In: J.R.Statist.Soc. B, vol. 45, no 1, 1983, pp. 47-50.

[11] F.Brickell and R.S.Clark, *Differentiable Manifolds, An Introduction,* 1970, Van Nostrand Reinhold Company.

[12] K.S.Chan and H.Tong, *A note on embedding a discrete parameter ARMA model in a continuous parameter ARMA model,* In: Journal of Time Series Analysis, vol. 8, no. 3, 1987, pp. 277-281

[13] R.Dahlhaus, *Fitting Time Series Models to Nonstationary Processes,,* 1993, University of Heidelberg, Germany.

[14] A.P.Dempster, N.M.Laird and D.B.Rubin, *Maximum Likelihood from Incomplete Data via the EM Algorithm,* In: Journal of the Royal Statistical Society B, vol. 39, 1977, pp. 1-38.

[15] J.E.Dennis Jr. and R.B.Schnabel, *Numerical Methods For Unconstrained Optimization And Nonlinear Equations,* 1983, Prentice-Hall.

[16] J.L.Doob, *The Brownian Movement and Stochastic Equations,* In: Annals of Mathematics, vol. 43, no. 2, 1942, pp. 351-369.

[17] J.C.Engwerda, *Regulation of Discrete Time-Varying Systems,* Ph.D.Thesis, 1988, Eindhoven, The Netherlands.

[18] J.C.Engwerda, A.C.M.Ran, A.L.Rijkeboer, *Necessary and Sufficient Conditions for the Existence of a Positive Definite Solution of the Matrix Equation* $X + A^*X^{-1}A = Q$, In: Linear Algebra and its Applications, vol. 186, 1993, pp. 255-275.

[19] L.Fahrmeir, *Posterior Mode Estimation by Extended Kalman Filtering for Multivariate Dynamic Generalized Linear Models,* In: Journal of the American Statistical Association, vol. 87, no. 418, 1992, pp. 501-508.

[20] L.Fahrmeir and H.Kaufmann, *On Kalman Filtering, Posterior Mode Estimation and Scoring in Dynamic Exponential Family Regression,* In: Metrika vol. 38, 1991, pp. 37-60.

[21] J.P.Florens and M.M.J.-M.Rolin, *Elements Of Bayesian Statistics,* 1990, Marcel Dekker, New York.

[22] T.C.Gard, *Introduction to Stochastic Differential Equations,* 1988, Marcel Dekker, New York.

[23] W.Gersch and T.Brotherton, *Estimation of Stationary Structural System Parameters from Non-Stationary Random Vibration Data: a Locally Stationary Model Method,* In: Journal of Sound and Vibration, vol. 81, no. 2, 1982, pp. 215-227.

[24] G.C.Goodwin and K.S.Sin, *Adaptive Filtering Prediction And Control,* 1984, Prentice-Hall.

[25] M.Green, *A Relative Error Bound for Balanced Stochastic Truncation,* In: IEEE Trans. Automat. Control., vol. AC-33, 1988, pp. 961-965.

[26] U.Grenander and M.Rosenblatt, *Statistical Analysis of Stationary Time Series,* 1956, Almqvist and Wiksell, Stockholm, Sweden.

[27] A.H.Gray and J.D.Markel, *Quantization and Bit Allocation in Speech Processing,* In: IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-24, no. 6, 1976, pp. 459-473.

[28] E.J.Hannan and M.Deistler, *The Statistical Theory of Linear Systems,* 1988, John Wiley & Sons.

[29] B.Hanzon, *Identifiability, Recursive Identification And Spaces Of Linear Dynamical Systems,* C.W.I. tracts no.63 and no.64, 1989, Centrum voor Wiskunde en Informatica, Amsterdam, The Netherlands.

[30] A.C.Harvey, *Forecasting, structural time series models and the Kalman filter,* Reprinted 1990, Cambridge University Press.

[31] S.W.He and J.G.Wang, *On embedding a discrete-parameter ARMA model in a continuous-parameter ARMA model,* In: Journal of Time Series Analysis, vol. 10, no. 4, 1989, pp. 315-323.

[32] S.Helgason, *Differential Geometry, Lie Groups, and Symmetric Spaces,* 1978, Academic Press.

[33] R.Hermann, *Differential Geometry and the Calculus of Variations,* 2nd edition 1977, Math Sci Press.

[34] P.Hougaard, *Parametrizations of Non-linear Models,* In: Journal of the Royal Statistical Society B, vol. 44, no. 2, 1982, pp.244-252.

[35] B.H.Jansen, *EEG Segmentation and Classification, An explorative study,* PhD thesis, 1979, Free University, Amsterdam, The Netherlands.

[36] A.H.Jazwinski, *Stochastic Processes And Filtering Theory,* 1970, Academic Press.

[37] C.F.Jeff Wu, *On the Convergence Properties of the EM Algorithm,* In: The Annals of Statistics, vol. 11, no.1, 1983, pp. 95-103.

[38] T.Kailath, *Lectures On Wiener And Kalman Filtering,* 1981, Springer-Verlag.

[39] L.V.Kantorovich and G.P.Akilov, *Functional Analysis,* 2nd. edition 1982, Pergamon Press.

[40] R.E.Kass, *Canonical Parametrizations and Zero Parameter-effects Curvature,* In: Journal of the Royal Statistical Society B, vol. 46, no. 1, 1984, pp. 86-92.

[41] G.Kitagawa *Changing Spectrum Estimation,* In: Journal of Sound and Vibration, vol.-89, no. 3, 1983, pp. 433-445.

[42] G.Kitagawa and W.Gersch, *A Smoothness Priors Time-Varying AR Coefficient Modelling of Non-stationary Covariance Time Series,* In: IEEE Transactions on Automatic Control, vol. AC-30, no. 1, 1985, pp. 48-56.

[43] G.Kitagawa and W.Gersch, *A Smoothness Priors Long AR Model Method for Spectral Estimation,* In: IEEE Transactions on Automatic Control, vol. AC-30, no. 1, 1985, pp. 57-65.

[44] S.Kobayashi and K.Nomizu, *Foundations Of Differential Geometry,* vol I, 1963, John Wiley & Sons.

[45] S.Kobayashi and K.Nomizu, *Foundations Of Differential Geometry,* vol. II, 1969, John Wiley & Sons.

[46] A.N.Kolmogorov and S.V.Fomin, *Introductory Real Analysis,* Revised English Edition, 1970, Prentice-Hall.

[47] R.J.Kugel, *Varying Parameter Models And Nonparametric Regression,* PhD thesis, 1990, University of Groningen, Groningen, The Netherlands.

[48] S.Kullback, *Information Theory and Statistics,* 1959, John Wiley & Sons, New York.

[49] H.J.Landau, *Maximum entropy and the moment problem,* In: Bull.A.M.S., vol. 16, nr. 1, 1987, pp. 47-77.

[50] L.Ljung, *System Identification: Theory for the User,* 1987, Prentice-Hall.

[51] L.Ljung *Optimal and ad hoc Adaptation Mechanisms,* In: ECC91 European Control Conference, Grenoble, France, 1991, July 2-5.

[52] L.Ljung and S.Gunnarsson, *Adaptation and Tracking in System Identification-A Survey,* In: Automatica, vol. 26, no. 1, 1990, pp. 7-21.

[53] L.Ljung and T.Söderström, *Theory and Practice of Recursive Identification,* 1983, The MIT Press, Cambridge, Massachusetts.

[54] H.B.Mann and A.Wald, *On the statistical treatment of linear stochastic difference equations,* In: Econometrica, vol. 11 no. 3 and 4, 1943, pp. 173-220

[55] P.McCullagh, *Tensor Methods in Statistics,* 1987, Chapman and Hall.

[56] P.McCullagh and J.A.Nelder, *Generalized Linear Models,* 2nd. edition, 1989, Chapman And Hall.

[57] G.Mélard and A.Herteleer-de Schutter, *Contributions to Evolutionary Spectral Theory,* In: Journal of Time Series Analysis, vol. 10, no. 1, 1988, pp. 41-63.

[58] J.Neveu, *Discrete-Parameter Martingales,* translated by T.P.Speed, 1975, North-Holland Publishing Company, Amsterdam, American Elsevier Publishing Company, New York.

[59] H.Nijmeijer and A.v.d.Schaft, *Nonlinear Dynamical Control Systems,* 1990, Springer-Verlag.

[60] I.V.Nikiforov, *Optimal Sequential Detection and Isolation of Changes in Stochastic Systems,* Publication Interne No.753, 1993, Institut de Recherche en Informatique et Systèmes Aléatoires, Rennes, France.

[61] R.L.M.Peeters, *System Identification Based on Riemannian Geometry: Theory and Algorithms,* Ph.D.Thesis, Tinbergen Institute Research Series, No. 64, 1994, Amsterdam, The Netherlands.

[62] R.L.M.Peeters and B.Hanzon, *The Riemannian Interpretation of Gauss-Newton and Scoring, with Application to System Identification,* In: Research Memorandum 1992-22, Free University, Amsterdam, The Netherlands.

[63] R.L.M.Peeters and B.Hanzon, *A Balanced Canonical Form for Discrete-Time Stable All-Pass Systems,* to appear in Proceedings MTNS, 1993, Regensburg, Germany.

[64] M.Pourahmadi, *Stationarity of the solution of* $X_t = A_t X_{t-1} + \epsilon_t$ *and analysis of non-gaussian dependent random variables ,* In: Journal of Time Series Analysis, vol. 9,no.3, 1988, pp. 225-239.

[65] M.B.Priestley, *Design Relations for Non-stationary Processes,* In: Journal of the Royal Statistical Society B, vol. 28, 1966, pp. 228-240.

[66] M.B.Priestley, *Evolutionary Spectra and Non-stationary Processes,* In: Journal of the Royal Statistical Society B, vol. 27, 1965, pp. 204-237.

[67] M.B.Priestley, *Spectral Analyis and Time Series,* Vol 1,2, 1981, Academic Press, London.

[68] C.R.Rao, *Information and accuracy attainable in the estimation of statistical parameters,* In: Bull. Calcutta. Math. Soc., vol. 37, 1945, pp. 81-91.

[69] N.Ravishanker, E.L.Melnick and C-L. Tsai, *Differential Geometry of ARMA models,* In: Journal of Time Series Analysis, vol. 11, no. 3, 1990, pp. 259-274.

[70] Yu.A.Rozanov, *Stationary Random Processes,* translation A.Feinstein, 1967, Holden-Day, San Francisco.

[71] I.Schur, *Über Potenzreihen, die im Innern des Einheitskreises beschränkt sind,* In: Journal für reine und angewandte Mathematik, vol. 147, 1917, pp.205-232.

[72] M.Vaeth, *On the Use of Wald's Test in Exponential Families,* In: International Statistical Review, vol. 53, no. 2, 1985, pp. 199-214.

[73] A.-J. van der Veen, *Time-Varying Systems, Theory and Computational Modeling. Realization, Approximation, and Factorization,* Ph.D.Thesis, 1993, Delft, The Netherlands.

[74] D. de Vries, *Identification of Systems with Time-Varying Dynamics,* 1992, Report, Tilburg University, Tilburg, The Netherlands.

[75] M.W.Watson and R.F.Engle, *Alternative Algorithms for the Estimation of Dynamic Factor, Mimic and Varying Coefficient Regression Models,* In: Journal of Econometrics, vol. 23, 1983, pp. 385-400, North-Holland.

[76] N.Wax, editor of *Selected papers on: Noise And Stochastic Processes,* 1954, Dover Publications, New York

[77] E.T. Whittaker, *On a new method of graduation,* In: Proc. Edinborough Math.Soc., vol. 41, 1923, pp. 63-75.

[78] N.Wiener, *Time Series,* 5th. printing 1977, The M.I.T. Press, Cambridge, Massachumetts.

[79] S.Wolfram, *Mathematica: A System for Doing Mathematics by Computer,* 1988, Addison-Wesley Publishing Company.

[80] R.Wu, *Stochastic differential equations,* 1985, Pitman Publishing Limited, London.

[81] P.C.Young,*Recursive Estimation and Time-Series Analysis* - An Introduction, 1984, Springer-Verlag.

[82] P.C.Young, C.N.Ng, *Recursive Estimation and Forecasting of Non-stationary Time Series,* In: Journal of Forecasting, vol. 9, 1990, pp. 173-204.

[83] P.C.Young, C.N.Ng, K.Lane and D.Parker, *Recursive Forecasting, Smoothing and Seasonal Adjustment of Non-Stationary Environmental Data,* In: Journal of Forecasting, vol. 10, 1991, pp. 57-89.

[84] P.C.Young, C.Ng, and P.Armitage, *A Systems Approach to Recursive Economic Forecasting and Seasonal Adjustment,* In: Computers Math. Applic, vol. 18, no. 6/7, 1989, pp. 481-501.

Index

Samenvatting

In dit proefschrift worden alternatieve modellen behandeld voor coëfficiënten van een tijdsveranderlijk autoregressief proces. Gebruikelijk is een model waarbij de coëfficiënten $\vec{a}_t$ zich als een (normaal verdeelde) random walk door $\mathbb{R}^n$ bewegen. Met andere woorden, bij het gebruikelijke model gaat men uit van een a priori normale kansverdeling voor de coëfficiënten $\vec{a}_0, \cdots, \vec{a}_T$ gelijk aan die van een random walk. De kansverdeling a posteriori (gegeven alle data $y_0, \cdots, y_T$) van deze coëfficiënten is dan ook normaal en kan met behulp van Kalman Filtering/Smoothing worden berekend uit $y_0, \cdots, y_T$. Men verkrijgt een random-walk-model voor de coëfficiënten door bijvoorbeeld een *lineaire* overgangsvergelijking (transitievergelijking) van de volgende vorm te geven:

$$\vec{a}_{t+1} = \vec{a}_t + \tau_t \lambda_t, \qquad (\lambda_t)_{t\in\mathbf{Z}} \text{ normale witte ruis met variantie gelijk aan de identiteit.}$$

Deze overgangsvergelijking impliceert dat de a priori conditionele kansverdeling van $\vec{a}_{t+1}$, gegeven het verleden $\vec{a}_0, \cdots, \vec{a}_t$, normaal is met verwachting $\vec{a}_t$ en variantie $\tau_t^2 I$. Merk op dat deze variantie niet afhangt van waar $\vec{a}_t$ ligt in de coëfficiëntenruimte. Het is een mooie en elegante eigenschap van het random-walk-model dat een sprong van $\vec{a}_t$ naar $\vec{a}_{t+1}$ met nominale spronggrootte

$$||\vec{a}_{t+1} - \vec{a}_t|| = \sqrt{\sum_{j=1}^{n} (\vec{a}_{t+1} - \vec{a}_t)_j^2} \tag{.41}$$

groter dan bijvoorbeeld 0.2 *overal in de coëfficiëntenruimte dezelfde kans heeft, ongeacht de waarde van* $\vec{a}_t$. Deze eigenschap van het random-walk-model is alleen mooi en elegant als we de manier van sprongmeting in vgl. (.41) a priori aanvaarden. Er is echter alle reden om die spronggrootte soms op andere wijzen te berekenen en ten opzichte van die andere sprongmaat heeft het random-walk-model dit natuurlijke kenmerk niet.
We geven een voorbeeld. We beschouwen een tijdsinvariant AR(1)-proces $(y_t)_{t\in\mathbf{Z}}$ dat op zeker tijdstip $t = k$ verandert van coëfficiënt. We kunnen dit als volgt in een model gieten:

$$a_{t+1} = a_t + \tau_t \lambda_t \tag{.42}$$
$$y_t = a_t y_{t-1} + \sigma \epsilon_t. \tag{.43}$$

Ook hier is de systeemruis $(\lambda_t)_{t\in\mathbf{Z}}$ normaal verdeelde witte ruis met variantie gelijk aan 1, evenals en onafhankelijk hiervan, de meet- of waarnemingsruis $(\epsilon_t)_{t\in\mathbf{Z}}$. De overgangshyperparameters τ_t zijn gelijk nul voor alle $t \neq k$. Een sprong met nominale spronggrootte tussen 0.04 en 0.05 op tijdstip $t = k$ krijgt van het model dezelfde waarschijnlijkheid in het geval dat $a_0 = a_1 = \cdots = a_k = 0.1$ als in het geval dat $a_0 = a_1 = \cdots = a_k = 0.9$. Toch zijn beide sprongen van geheel verschillende aard. Dit blijkt bijvoorbeeld als men de (asymptotische) Kullback-Leibler divergentie berekent in beide gevallen. De asymptotische Kullback-Leibler divergentie is een veel gebruikte maat voor de afstand tussen stationaire AR processen. De Kullback-Leibler divergentie van het AR(1)-proces $y_t = 0.9 y_{t-1} + \sigma\epsilon_t$ met het proces $y_t = 0.95 y_{t-1} + \sigma\epsilon_t$ is veel groter dan de Kullback-Leibler divergentie tussen de AR(1)-processen gegeven door respectievelijk $y_t = 0.1 y_{t-1} + \sigma\epsilon_t$ en

$y_t = 0.05y_{t-1} + \sigma\epsilon_t$.
In dit proefschrift hebben we getracht modellen te formuleren die de hierboven vermelde mooie, elegante eigenschap hebben, dat de spronggrootte onafhankelijk is van de positie in de coëfficiëntenruimte, maar waarbij het meten van die spronggrootte niet noodzakelijk geschiedt als in vgl. (.41). Onze theorie maakt het mogelijk a priori de wijze van sprongmeting vast te leggen door middel van de keuze van een metriek op de coëfficiëntenruimte.
Verder vergroten wij ook de keuzemogelijkheden voor de coëfficiëntenruimte: we gebruiken niet meer automatisch de hele $\mathbb{R}^n$ maar bijvoorbeeld alleen het deel van $\mathbb{R}^n$ waarvan de elementen corresponderen met coëfficiënten van *stationaire* tijdsinvariante AR(n)-processen, het "stabiliteitsgebied". In het algemeen zal de coëfficiëntenruimte een willekeurige differentieerbare variëteit zijn. Wij beschouwen metrieken op de coëfficiëntenruimte zodat deze coëfficiëntenruimte met de ingevoerde metriek een zogenaamde Riemann-variëteit wordt. In hoofdstuk drie veralgemenen we het begrip van de "random walk" in $\mathbb{R}^n$ tot een proces op een Riemann-variëteit op een dusdanige wijze dat de mooie, elegante eigenschap, dat de spronggrootte onafhankelijk is van de positie in coëfficiëntenruimte, gehandhaafd blijft, maar nu ten opzichte van de metriek van de Riemann-variëteit. Een dergelijk proces zullen wij een "veelhoek-proces met onafhankelijke variatie in richting" noemen. Zo'n proces zal het basismodel zijn voor het coëfficiëntenproces $(\vec{a}_t)_{t\geq 0}$ van een tijdsveranderlijk autoregressief proces $(y_t)_{t\geq 0}$. Dit basismodel noemen wij het Speciale Model.
Behalve de random walk wordt ook wel de vloeiend geïntegreerde random walk als model voor tijdsveranderlijke coëfficiënten van een autoregressief proces genomen. Dan is de overgangsvergelijking weer *lineair* en van de vorm

$$\vec{a}_{t+1} = \vec{a}_t + \tau_t d_t \qquad (.44)$$

$$d_t = \mu d_{t-1} + \lambda_t \qquad (.45)$$

waarbij de hyperparameter μ een element is van $[0,1)$ (en $(\lambda_t)_{t\in\mathbb{Z}}$ weer normaal verdeelde witte ruis is met variantie gelijk aan de identiteit). Ook de vloeiend geïntegreerde random walk kan op dezelfde manier als de gewone random walk worden gegeneralizeerd tot een veelhoek-proces op een Riemann-variëteit met onafhankelijke variatie in richting. Het gebruik van een dergelijk proces om het coëfficiëntenproces te modelleren leidt tot wat wij noemen het Algemene Model.
Als wij op de coëfficiëntenruimte een Euclidische metriek kiezen, dan blijft de overgangsvergelijking van het Speciale en Algemene Model lineair en voeren de coëfficiënten een random walk of een vloeiend geïntegreerde random walk uit. Dan blijft alles bij het oude model. Maar als we een andere metriek dan de Euclidische kiezen, dan is de overgangsvergelijking van een Speciaal of Algemeen Model niet lineair meer. We kunnen echter andere coördinaten invoeren in de coëfficiëntenruimte, en dan kijken hoe de overgangsvergelijking in die coördinaten eruit ziet. In hoofdstuk drie worden nodige en voldoende voorwaarden gegeven voor het Algemene en het Speciale Model zodat men coördinaten κ in de coëfficiëntenruimte kan invoeren op een dusdanige manier dat de overgangsvergelijking lineair wordt, d.w.z. de betreffende coördinaten κ_t a priori een gewone of vloeiend geïntegreerde random walk uitvoeren. Die voorwaarden hebben met afwezigheid van kromming in de coëfficiëntenruimte te maken; daarom

noemen wij Algemene en Speciale Modellen *nul-krommings-modellen* als het mogelijk is opnieuw te parametrizeren zodat de overgangsvergelijking lineair wordt. De bijbehorende vergelijking voor de waarnemingen,

$$y_t = \psi_1(\kappa_t)y_{t-1} + \cdots + \psi_n(\kappa_t)y_{t-n} + \sigma\epsilon_t$$

bij een dergelijke parametrizering is nu niet lineair meer in κ_t. De a-posteriori-kansverdeling van $\kappa_0, \cdots, \kappa_T$, gegeven $y_0, \cdots, y_T$, is dan ook niet normaal meer. In hoofdstuk zes wordt aangetoond dat voor nul-krommings-modellen het, onder zekere voorwaarden, toch mogelijk is belangrijke aspecten van deze a-posteriori-kansverdeling, zoals 'mode' (of maximum) en tweede afgeleide van log-a-posteriori-kansdichtheid, te achterhalen door middel van een benaderingsmethode gebaseerd op Gauss-Newton-iteraties. In dat hoofdstuk wordt ook het verband gelegd tussen deze methode en Extended Kalman Filtering/Smoothing, waarbij wij voortbouwen op werk van Fahrmeir en Kaufmann. We laten daar ook zien, dat voor toepassing van die Gauss-Newton-iteraties het niet nodig is de analytische formules van de parametrizering te kennen waarop de overgangsvergelijking lineair wordt; we geven een benaderingsprocedure voor deze parametrizering, waarbij alleen kenmerkende grootheden voor de metriek als Christoffel- en krommings-symbolen hoeven te worden uitgerekend.
Een belangrijk voorbeeld van een nul-krommings-model is het zogenaamde Geodetische Model, dat bestaat voor elke Riemann-metriek. Hierbij is de a-priori veronderstelling dat de coëfficiënten zich vanaf het begin bevinden op één geodeet, d.w.z. een rechte lijn in de zin van de Riemann-meetkunde gegeven door de metriek op de coëfficiëntenruimte. De vooronderstelling van het model kan dan bijvoorbeeld zijn, dat die geodeet een van de geodeten is door een bekend punt in de coëfficiëntenruimte, waarbij de richting van de geodeet een hyperparameter ζ is die uit de waarnemingen geschat moet worden. Ook het waarnemingsruisniveau σ^2 en de hyperparameters in de overgangsvergelijking, zoals de τ_t en de μ, moeten we halen uit de waarnemingen. Het criterium voor optimaliteit van deze hyperparameters is gebaseerd op maximum-likelihood. Wij hebben enige ad-hoc-methoden opgesteld om deze hyperparameters te bepalen, gebaseerd op Gauss-Newton-iteraties, Expectation-Maximization en een eenvoudige methode waarbij de parameterruimte wordt afgetast. De eerste en laatste methoden zijn met goed resultaat toegepast op een groot aantal gesimuleerde tijdreeksen, voornamelijk tijdens een stimulerend verblijf aan Professor Young's Centre for Research on Environmental Systems aan de universiteit van Lancaster in Engeland. Een theoretische analyse van de eigenschappen van de gepresenteerde methoden om de hyperparameters te bepalen (bijvoorbeeld asymptotische verdelingen van schatters van de hyperparameters) blijft echter buiten beschouwing in dit proefschrift. Een grondslag voor een dergelijke theoretische analyse kan echter asymptotiek zijn. Hiermee snijden we een ander hoofdonderwerp aan dat in dit proefschrift aan de orde komt.
In hoofdstuk vier beschouwen we het moeilijke vraagstuk wat onder asymptotiek voor *tijdsveranderlijke* autoregressieve processen verstaan moet worden. Asymptotiek van tijdsinvariante, stabiele autoregressieve processen is een wel bekend onderwerp in de theorie van stationaire processen. Wij volgen en bouwen voort op de ideeën van Dahlhaus om het begrip van asymptotiek in deze theorie te veralgemenen tot een dergelijk begrip voor *rijen* tijdsveranderlijke autoregressieve processen. Wij laten enige mooie eigen-

schappen van deze generalisatie zien; bijvoorbeeld dat de equivalentie van stabiliteit met het hebben van een spectrale dichtheid in het geval van tijdsinvariante autoregressieve processen gehandhaafd blijft in het algemenere geval van rijen tijdsveranderlijke autoregressieve processen; de spectrale dichtheid is dan eveneens tijdsveranderlijk. Wij laten zien dat de coëfficiënten van elk proces uit zo'n rij tijdsveranderlijke autoregressieve processen stochastisch gemodelleerd kunnen zijn, en wel in overeenstemming met het Algemene Model. De tijdsveranderlijke spectrale dichtheid en ook de stabiliteitsvoorwaarde zelf is dan conditioneel op het coëfficiëntenproces. De stabiliteitsvoorwaarde blijkt geformuleerd te kunnen worden in de vorm van de waarden van een stoptijd. Tevens laten we zien dat voor Algemene Modellen gebaseerd op een zogenaamde volledige metriek op het stabiliteitsgebied deze stabiliteitsvoorwaarde automatisch gerealiseerd is.
Het idee van de asymptotische, tijdsveranderlijke spectrale dichtheid bewijst zijn nut, doordat we in staat zijn een mooie eigenschap af te leiden voor Geodetische Modellen gebaseerd op de asymptotische Fisher metriek. De boven vermelde asymptotische Kullback-Leibler divergentie tussen tijdsinvariante AR(n)-processen is infinitesimaal (d.w.z. bij naar nul gaande divergentie) gelijk aan het kwadraat van de Riemann-afstand gebaseerd op deze metriek. In hoofdstuk 6, sectie 6.1.4, laten we zien dat in zekere zin in deze modellen, gebaseerd op de asymptotische Fisher metriek, de vooronderstelling besloten ligt, dat asymptotisch de bijdragen van alle waarnemingen aan de informatie over de coördinaten van de coëfficiënten gelijk zijn. We brengen deze eigenschap in verband met de waarneembaarheidsgramiaan en we brengen de boven vermelde eigenschap van "onafhankelijke variatie in richting" van het veelhoek-coëfficiëntenproces in verband met de bestuurbaarheidsgramiaan. Het een en ander wordt iets nader uitgewerkt in hoofdstuk 7 (de conclusies). Ook het "enkele-sprong"-voorbeeld aan het begin van de Samenvatting toont aan dat het interessant is naar deze asymptotische Fisher metriek te kijken als zinvol alternatief voor de gewone Euclidische metriek, waarop de gebruikelijke random-walk modellen voor het coëfficiëntenproces gebaseerd zijn. Een uitgebreide studie van deze metriek wordt ondernomen in hoofdstuk vijf. Dit hoofdstuk laat zich grotendeels lezen los van de rest van het proefschrift en is ook van belang als men alleen de modelverzameling van stationaire tijdsinvariante AR(n)-processen wil bestuderen, zonder het oogmerk tijdsveranderlijke processen aan te pakken. Bijvoorbeeld is zo'n studie van belang als men zoekt naar parametrizaties van deze modelverzameling (of van stukken ervan) zodat maximum-likelihoodschatters van de coördinaten $(\kappa_1, \cdots, \kappa_n)$ van de coëfficiënt $\vec{a} = (a_1, \cdots, a_n)$ asymptotisch een variantie hebben die niet afhangt van de verwachting van die schatter. Zulke parametrizaties noemt men wel "covariantie stabilizerend". In hoofdstuk vijf tonen we aan dat er niet een parametrizatie bestaat van de hele modelverzameling die de covariantie stabilizeert, maar wel zo'n parametrizatie voor sommige stukken ervan. Wij laten zien dat de Riemann-afstand gebaseerd op de asymptotische Fisher metriek op de modelverzameling van stationaire AR(n)-processen met gelijk ruisniveau eindig is, en dat de metriek niet volledig is. Onze studie blijkt een vruchtbaar vervolg op de theorie van de Schurparameters. We slagen erin de Riemann-meetkunde gebaseerd op de asymptotische Fisher metriek, volledig te beschrijven (met analytische formules voor de geodeten) in het geval van de modelverzameling van AR(2) processen. Dit resultaat heeft een praktische toepassing, wanneer we tijdsveranderlijke AR(2) processen volgens het Algemene Model, gebaseerd op de asymptotische Fisher

metriek willen simuleren; dat kan nu exact.
In hetzelfde hoofdstuk geven we ook een voorbeeld van een volledige metriek op de modelverzameling van stationaire AR(n)-processen. Tot slot merken we op dat we in het hele proefschrift voortdurend ook rekening houden met een stochastische modellering van een tijdsveranderlijk niveau van de waarnemingsruis en we geven enige ad-hoc methoden om dit tijdsveranderlijke niveau te identificeren.
Het hoofddoel van ons onderzoek is geweest door de keuze van andere metrieken dan de Euclidische het mogelijk te maken in vele gevallen tot een preciezere beschrijving van de dynamiek van een tijdsveranderlijk autoregressief proces te komen zonder het aantal hyperparameters te vergroten. Een preciezere beschrijving van de dynamiek maakt het natuurlijk mogelijk betere voorspellingen te doen over de toekomst van het beschouwde proces.